Information diffusion and opinion dynamics in social networks

Julio Cesar Louzada Pinto

THÈSE DE DOCTORAT DE
TELECOM SUDPARIS

Spécialité

Informatique

École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Julio Cesar LOUZADA PINTO

Pour obtenir le grade de

DOCTEUR DE TELECOM SUDPARIS

Sujet de la thèse :

Dissémination de l'information et dynamique des opinions dans les réseaux sociaux

Soutenue le 14 janvier 2016

Devant le jury composé de :

M. Jérémie Jakubowicz	Encadrant
M. Eitan Altman	Directeur de thèse
M. Tijani Chahed	Directeur de thèse
M. Rachid El-Azouzi	Rapporteur
Mme Nidhi Hegde	Rapporteur
M. Tamer Başar	Examinateur
M. Guy Pujolle	Examinateur

N° NNT: 2016TELE0001

UNIVERSITÉ PIERRE ET MARIE CURIE

UPMC

LABORATOIRE SAMOVAR

Thesis presented by

Júlio Cesar Louzada Pinto

as requirement for the degree of

Doctor of Philosophy
in
Computer Science

Title:

Information diffusion and opinion dynamics in social networks

Defended January 14^{th} 2016 in front of the jury composed by:

M.	**Rachid El-Azouzi**	**Referee**
Mme	**Nidhi Hegde**	**Referee**
M.	**Tamer Başar**	**Examiner**
M.	**Guy Pujolle**	**Examiner**
M.	**Jérémie Jakubowicz**	**Advisor**
M.	**Eitan Altman**	**Advisor**
M.	**Tijani Chahed**	**Advisor**

Abstract

Information diffusion explores the paths taken by information being transmitted through a social network in order to understand and model the relationships between users in such network, leading to a better comprehension of human relations and dynamics.

Although primarily concerned with theoretical, psychological and sociological aspects of social networks, information diffusion models also serve as basis for several real-life applications of social networks analysis, such as influence maximization, link prediction, discovery of influential nodes, community detection, trend detection, etc.

This thesis is thus based on both sides of information diffusion: first by developing mathematical theories and models to study the relationships between people and information, and in a second time by creating tools to better exploit the hidden patterns in these relationships.

The theoretical tools developed here are opinion dynamics models and information diffusion models, where we study the information flow from users in social networks, and the practical tools developed here are a novel community detection algorithm and a novel trend detection algorithm.

We start by introducing an discrete-time opinion dynamics model in which agents interact with each other about several distinct opinions/contents. In our framework, agents do not exchange all their opinions with each other, they communicate about randomly chosen opinions at each time. Our model uses scores to take this dynamics into account: each agent maintains a list of scores for each opinion held. Opinions are selected according to a softmax choice function based on their scores (the higher a score, the more likely an opinion is to be expressed) and then transmitted to neighbors. Once an agent receives an opinion it gives it more credit, i.e., a higher score to this opinion. We show, using stochastic approximation algorithms, that under mild assumptions the opinion dynamics algorithm converges as time increases, whose behavior is ruled by how users choose the opinions to broadcast at each time.

We develop next a practical algorithm which is a direct application of a particular instance of this opinion dynamics model. When agents broadcast the content they appreciate the most, communities are formed in the social network, where these communities are groups of users that broadcast the same piece of information. This community detection algorithm, which is distributed by nature, has the remarkable property that the discovered communities can be studied from a solid mathematical standpoint. In addition to the theoretical advantage over heuristic community detection methods, the presented algorithm is able to accommodate weighted and directed networks; parametric and nonparametric versions; and the discovery of overlapping communities as a byproduct with no mathematical overhead.

In a second part, we define a general Hawkes-based framework to model information diffusion in social networks. The proposed framework takes into consideration not only the hidden interactions between users but as well the interactions between contents and social networks, and can also accommodate dynamic social networks and various temporal effects of the diffusion, which provides a

complete analysis of the hidden influences in social networks. This framework can be combined with topic modeling, for which modified collapsed Gibbs sampling and variational Bayes techniques are derived. We provide an estimation algorithm based on nonnegative tensor factorization techniques, which together with a dimensionality reduction argument are able to discover, in addition, the latent community structure of the social network.

Finally, we apply one instance of the previous information diffusion framework and develop stochastic control techniques for near unstable Hawkes processes, creating a trend detection algorithm, designed to find trendy topics being disseminated in a social network. We assume that the broadcasts of messages in the social network is governed by a self-exciting point process, namely a Hawkes process, which takes into consideration the real broadcasting times of messages and the interaction between users and topics. We formally define trendiness and derive trend indices for each topic being disseminated in the social network. These indices take into consideration the time between the detection and the message broadcasts, the distance between the real broadcast intensity and the maximum expected broadcast intensity, and the social network topology. The proposed trend detection algorithm is simple and uses stochastic control techniques in order calculate the trend indices. It is also fast and aggregates all the information of the broadcasts into a simple one-dimensional process, thus reducing its complexity and the quantity of necessary data to the detection. The advantage of this trend detection algorithm is that, to the best of our knowledge, this is the first trend detection algorithm that is based solely on the individual performances of topics, i.e., a topic may have a relatively small number of broadcasts and still be considered trendy. The trendiness is thus, in our case, an expectation of an increase in broadcasts, not a comparison measure with other topics.

Keywords: Opinion dynamics, stochastic approximation algorithms, community detection, information diffusion, Hawkes processes, trend detection, stochastic control.

Contents

Acknowledgements

"give thanks in all circumstances, for this is God's will for you in Christ Jesus."

— I Thessalonians*, 5:18*

First and foremost I would like to thank God for guiding me during this Ph.D. His thoughts guided my thoughts when researching, His hand guided my hands when writing this manuscript. On the darkest hours He was there for me, giving me the strength so much needed to finish it. Without Him, this thesis would never exist.

I do not have the words to express myself about how much I want to thank my thesis advisor Tijani Chahed; there are so many things to be thankful for that I would have to write a whole thesis on the subject: his knowledge and resourcefulness are only second to his patience towards me; the long discussions and the pivoting during our research were paramount to shape the thesis as you now read it. The freedom bestowed upon me allowed a full enjoyment of every second of this work, and his trust in my abilities gave me the responsibility and self-awareness that failure or a half-baked thesis were not options.

I am also extremely thankful to Jérémie Jakubowicz. He is a true example of what a researcher and an applied mathematician should aspire to be. Even though he is not officially my advisor, I have been taken under his wings during my Ph.D. and received valuable advice about the world of research and mathematics. Despite his tight schedule, he has always shown an amazing willingness to work, a curiosity and rigor worth of a world-class mathematician, and in a handful of moments gave me very inspirational and (even more) correctional speeches - I am pretty sure that he was testing these speeches on me to use on his children later in life, and I am thankful for that! I take tremendous pleasure, joy and pride in calling Jérémie and Tijani not only my co-workers, but also my friends.

A special thanks also goes to Eitan Altman and the Avignon team. The distance between Avignon and Paris was not enough to hinder our efforts to collaborate and work together. Eitan's multitude of ideas, availability and enthusiasm were much appreciated and very fruitful works emerged nicely from our time - albeit small - together. A most warmful thanks to Alexandre Reiffers, who helped in every way he could: providing datasets, discussions, brainstorming, a house to stay and very, very hard-core jokes.

I would like to thank as well the members of the jury - specially the external readers - for their brilliant comments and suggestions, and for letting my defense be an enjoyable moment for everyone.

I could nor would never leave without words to my friends over both sides of the Atlantic: during the moments that the family was not there to support me, you took her place. I have met amazing and extraordinary people during my lifetime, and those that I have met after arriving in France are

not out of that list. They have shown me a different way of life, of thinking, and that everything in life is worthy if you pour your soul into it; that job is not only a place to go every morning (for those that are not or never were researchers), but a place to be enjoyed and looked forward to go to. I would not dare to write down any name here since by doing so I will certainly leave a lot of important ones out as well. For that and much more, thank you from the bottom of my heart.

It goes without saying - but I will say anyway - that a very special and loving thanks goes to my family: my mother, my father and my sister (with of course every other one that knows me well enough to place himself inside my family circle). They were - and still are - one of the most important things in my life, and without them I would be utterly and completely lost from the beginning. They have molded my character and my ideas from my mother's womb, and have given and shown me not only knowledge, charity, principle, God, but also love - tons and tons of love. I can say, without shadow of doubt, that I am the man, husband and researcher that you all know today thanks to them, and I will be forever in their debt.

I could not finish any acknowledgement without saying a few words about the most important person in my life: my wife Karina. She is the reason that I wake up every day with a smile in my face, and go to sleep every day with a bigger smile in my face. She is my soul mate and my one true love, responsible for making me become a better man, better husband and if God wishes one day a great father. When the family was far away, she was my earthly pillar. She was the voice through which God guided me during my worst moments, and the light that pulled me out of darkness so many times that I lost count. I can honestly say that she made me the happiest man alive by marrying me, and keeps me still by staying next to me every day. Without her, this thesis would have been not only much more difficult, but quite possibly nonexistent. As she said to me once: thank you very much for making my life more beautiful and sweeter!

Introduction

"We're all social creatures."
— Social Media Monthly Magazine

In an era where information travels fast and far away, people become more interconnected and aware of their surroundings, their planet, and other cultures. This phenomenon facilitates the formation of public opinions, mass movements, political polarization, and at the same time, the creation of niches where people assemble around some specific interest or quirk, such as skydiving, bodybuilding, cosplay, etc.

The one responsible for this informational advance was, without shadow of doubt, the *Internet*. It was entirely fundamental for the development of this news-centered paradigm, which is a cyber environment composed of social networks, social media sites, web blogs, and any other means of digital communication.

As we became capable of storing more and more data and performing faster calculations, social networks and social sites such as Google, Facebook, Twitter and Amazon became data centers where everyone willing to exploit this huge amount of information circles around.

As humans connect in increasing rates, their relationships, stories, experiences and knowledge interact, which creates a multitude of crowd behaviors that may lead to a better understanding of human relationships. As a consequence, social networks became one of the major research themes during these last years, as a source of never-ending untapped information serving researchers in many areas of science: psychology, sociology, advertising, statistics, physics, computer science, etc.

A good example of crowd behavior with concrete applications is the *wisdom of crowds*: people weight what others think and suggest, what most of their acquaintances do, where do they go to have a drink, eat something exotic, or simply enjoy life. This social knowledge is the basis of the social network *Trip Advisor*, which is nowadays stamped in most of the entertainment places all over the world.

A famous quote by W. Edwards Deming: *"In God we trust; all others bring data"* illustrates this accomplished almighty power of data and thus, also illustrates the ongoing role of social networks as one of the shapers and suppliers of data and knowledge about human interactions in the modern society. Moreover, the economic repercussions of this new business intelligence and business analytics model create the perfect incentive for data-driven and technology-driven companies, who thrive in this flicker environment and become the major players in innovation. This synergy between data and business generates hence a rising framework for the economical exploitation of this "big data", and as consequence, motivates more academic and applied research.

This ongoing trend, which generates an enormous flow of business and sheds some light into the nature of human behavior and human relationships, needs to be studied and understood; and for that matter a solid and extended mathematical set of tools to comprehend and exploit these new

paradigm of information is paramount.

That is precisely the framework of this thesis, which aims to

(i) develop mathematical methods to study these intricate relationships between people and information, and

(ii) create tools to better exploit the hidden patterns in these relationships.

This task is performed in two steps. The first one is theoretical and uses tools from behavioral dynamics - opinion dynamics models and information diffusion models. The second step is from a more practical and applied point of view, and uses the developed theoretical models in order to derive a novel community detection algorithm and a novel trend detection algorithm.

This thesis is thus divided into two parts, each one with a theoretical chapter and a practical application stemming from the developed theory. The first part of this thesis - composed by chapters 1 and 2 - develops a theoretical opinion dynamics algorithm for information diffusion in social networks, and a subsequent and more practical community detection algorithm. The second part - composed by chapters 3 and 4 - develops a theoretical information diffusion framework based on point processes, with a subsequent and more practical trend detection algorithm.

Specifically, chapter 1 develops an opinion dynamics model that takes into account multiple contents being broadcasted in a social network, and stochastic interactions between users. Broadcast means that users of the social network post the contents in their walls, which their followers are able to see/receive.

Chapter 2 is a direct application of a particular instance of this opinion dynamics model. When agents broadcast the content they appreciate the most, communities may be formed in the social network, where these communities are groups of users that broadcast the same piece of information. We have developed thus a community detection algorithm, which is distributed by nature, and has the property that the discovered communities can be studied from a solid mathematical standpoint, which is a valuable feature for a community detection algorithm.

Chapter 3 develops a theoretical framework for information diffusion based on self-exciting point processes - Hawkes processes [140, 208]. This framework is versatile and accommodates several variants, all of them possessing the same basic estimation procedure. The proposed framework discovers hidden influences between users and contents, and can be coupled with topic models in order to have a more data-driven approach.

Chapter 4 is a direct application of one instance of the previous information diffusion framework, where we develop stochastic control techniques for near unstable Hawkes processes, creating a trend detection algorithm. These techniques come from optimal stopping theory of Itô diffusions (Brownian diffusions), and define trend indices for each topic being disseminated. The advantage of this trend detection algorithm is that, to the best of our knowledge, this is the first trend detection algorithm that is based solely on the individual performances of topics, i.e., a topic may have a relatively small number of broadcasts and still be considered trendy. The trendiness is thus, in our case, an expectation of an increase in broadcasts, not an absolute comparison measure with other topics.

The remainder of this introduction discusses in more details the academic literature relative to each chapter of this thesis, providing a better comprehension of the tools used during the course of this work.

0.1 Opinion dynamics and community detection

As already mentioned, the first part of this thesis is dedicated to an *opinion dynamics* model and a subsequent *community detection* algorithm employing ideas from one of its particular cases. We provide now a general review of the literature about both subjects.

0.1.1 Opinion dynamics

Opinion dynamics is an active area of research among statistical physicists, mathematicians and computer scientists, started by psychologist French in [105], who created simple and intuitive models of interactions between people and information.

One of the first mathematical models in this area was developed by DeGroot in [83], where agents in a communication/trust graph possess scalar opinions about a subject, and at each discrete time step randomly selected agents perform a combination of their opinions with those of their neighbors, this combination being dictated by the weights of the trust graph. DeGroot shows in his work that if the graph satisfies some assumptions related to its connectivity - if it is for example strongly connected - then these individuals reach consensus, i.e., they converge to the same opinion.

Although DeGroot's model reflects in part the simple but correct idea that people tend to share and weight their opinions against peers, consensus is not always achieved in our society, even in small groups. In practice we do not know when a group of people will achieve a consensus, or will subdivide in small groups that share the same opinions, i.e., clusters of people.

These cluster formations are the main subject of a different family of opinion dynamics models called *bounded confidence models*, for which the two major representatives are the Krause-Hegselmann model [142] and the Weisbuch-Deffuant model [82]. Bounded confidence models' premise is that people are inclined to accept other opinions when they are sufficiently close to one's own opinion. Krause presents in [179] a continuous state model in which agents have scalar opinions and at each discrete time step every individual averages his opinion with those that are sufficiently similar to his, i.e., that reside in his so-called *acceptance neighbourhood*. Another example of a bounded confidence model is the Deffuant's model [82], which is a stochastic version of Krause's model where at each time step two random individuals meet and if their opinions are sufficiently close then they average their opinions, which consequently approaches each other.

Krause and Hegselmann [142] and Lorenz [210] prove that, under mild assumptions on the graph connectivity and the opinion thresholds, both the Krause-Hegselmann model and the Weisbuch-Deffuant model present clustering, i.e., smaller groups having the same intra-group opinions are formed, although the study of the most general cases can only be tackled by numerical simulations. In spite of all difficulties to derive precise analytical results predicting the behavior of these models, several extensions have been developed (see *e.g.* [89, 90, 91]).

As one may notice, there are as many opinion dynamics models as there are particular behaviors that need to be modeled and understood in social, economic and natural sciences. We discuss some families of opinion dynamics models that reproduce and explain different mechanisms of interaction between agents, be they people, animals, competing industries, etc.

The principal families of opinion dynamics models studied here are: *Interacting particle systems* and *Crowd behavior models*. These two families differ on the account that interacting particle systems deal mostly with discrete states for agents and simple rules of interaction between them, whereas crowd behavior models use basically differential equations in order to derive average non-trivial behavior for the system of agents.

These two families can be divided into several subfamilies, each one representative of an attempt

to model a particular feature of reality. For example, interacting particle systems can be further categorized into

- *Contact processes:* developed by Clifford and Sudbury in [68] for the study of two species competing for territory, each one possessing a binary state. Each competitor thus can change its state with a rate depending on the quantity of neighbors in each space, giving rise to a Markov process in the state space.

  Other examples of stochastic processes describing contact (between people in a society, or nodes in a graph) are [137], where the stochastic contact model is represented by a discrete state. Holley and Liggett study in [153] the ergodicity and stationary properties for this kind of stochastic contact process, providing a complete description of its mechanisms.

- *Voter models:* these models are the modern version of contact processes in the sense that they too are based on interactions between agents and their neighbors, in discrete or continuous time. As with the contact process, the behavior in these models depends greatly on the type of interactions between agents, which can be of any sort.

  For example, Cox and Durrett introduce in [70] the nonlinear voter model, having as particular example the threshold voter model of Liggett [204], which models the rate of an agent with a binary state space to change his state as a threshold function of his neighbors states, i.e., an agent has a nonzero probability of changing his state if he possesses sufficiently many neighbors in the other state. Lanchier and Neuhauser propose in [187] a (biased) voter model in a heterogeneous environment to investigate the effects of recurrent gene flow from transgenic crop to wild relatives, with a binary state space (individuals carrying the transgenic gene or the wild gene) model where agents are situated in the d-dimensional lattice, for which there exists a predefined set in which the voter dynamics is frozen, i.e., agents belonging to this set do not change their initial states (which is the transgenic state). Yildiz *et al.* develop in [311] a voter model with possible stubborn agents; in this model agents communicate via a network (which is coded by a directed graph) and possess a binary state space; if not stubborn, an agent revises his state at independent exponential random times; the presence of stubborn agents in this model can lead to either consensus, as in DeGroot's case, or disagreement, as shown in [311].

- *Majority opinion models:* originally developed by Galam [108] to model public debates and hierarchical votes in societies, they are based on the assumption that in a random group of people reaching a consensus, they end up with the majority's opinion.

  For example, Galam proposes in [109] a model with individuals possessing binary opinions (in favor or against war, preferring a political candidate over another, etc.) who engage in discussions about these opinions; at each time step, a subgroup of individuals is randomly selected (following a distribution that takes into consideration their environment) and is put to discussion, with each individual adopting the majority opinion at the end of the discussion period (when opinions are tied, agents maintain the same opinions); the author finds that in this dynamical model there exists a threshold value such that initial concentrations of the minority opinion above this threshold imply that every individual eventually adopts it.

  In a different kind of model - the classical Sznajd model [275] - individuals occupy binary states in a linear chain, which are updated in a sequential order using the following rule: if an individual shares the same state as his neighbor, then he and his neighbor propagate this state to their respective neighbors; however if they do not share the same state, then their

respective neighbors' neighbors adopt their states, i.e., if neighbors share the same state, they influence the other neighbors to adopt it, but if they do not share the same state, then they influence their neighbors' neighbors to do so. The Sznajd model presents some interesting properties [274], but its assumptions are a little unrealistic, so variants and extensions have been proposed (see *e.g.* [274, 81, 110]).

- *Statistical physics models:* At last, we discuss models stemming from statistical physics, which in many cases bear resemblance to beforementioned models such as the voter model. These models are based on a function measuring the energy of the interaction between agents in a system, depending on the configuration of their states. Thus, authors develop updating rules for the changes in agents' states taking into account this "social Hamiltonian".

  A pioneering example of such model is the theory of *social impact*, developed by Latané in [190] and further studied by Nowak *et al.* in [239]. Again, we have a group of individuals who are able to communicate following a network, with two possible "spins"/opinions, and two individual parameters for each agent: his persuasiveness and supportiveness, describing how he can influence and be influenced by others, respectively. These parameters are supposed to be random, and give rise to a Hamiltonian function measuring the impact that an individual receives from the environment; this Hamiltonian takes into consideration not only agents spins/opinions, persuasiveness, supportiveness but also their distance in this network and random fields expressing all sources other than individuals that can alter opinions (for example media sources).

  Due to the simplicity of this model (for example, it does not take into consideration memories of individuals), some extensions have been proposed [175, 155], such as one based on active Brownian particles [270, 269].

A different paradigm of opinion dynamics and social behavior is crowd behavior models. This type of models translate the natural behavior into mathematical equations of motion, be they for schools of fish, flocks of birds, pedestrian movements in streets, vehicular traffic, etc. Crowd behavior models can be further categorized into

- *Flocking models:* Models in this category represent the behavior of flocks of birds, schools of fish and other natural phenomena by motion equations.

  A celebrated flocking model is the Cucker and Smale flocking model, developed in [72] to study the behavior of flocks of birds in mid-air, which is then given by a dynamical system accounting for the difference in distance and velocity between the birds in the flock; the authors are able to prove the existence of two very distinct regimes: a regime where the convergence of the flock is guaranteed, i.e., the flock remains together, and a regime where its convergence is not guaranteed.

  A stochastic version of Cucker and Smale flocking model is the Vicsek model [73] of *self-propelled particles*, where particles move in a square surface with periodic boundary conditions, and at each time step a particle position is updated following a law that takes into consideration the average behavior of its neighborhood and a uniformly randomly distributed noise.

- *Pedestrian behavior models:* Pedestrian behavior is studied empirically since the 1950's [134], with a pioneer model proposed in [148] conjecturing that pedestrian flows behave like gases or fluids, thus described by classical Navier-Stokes equations.

A celebrated example of this family is the *social force model*, developed by Helbing and coworkers [146, 147]. They model pedestrians as particles following Newtonian equations of motion containing three parts: a personal force that drives the pedestrian velocity, a social force given by interactions with other pedestrians and interactions with barriers, and a noise term responsible for the pedestrians non-predictable behavior. This simple model reflects realistic scenarios, such as the formation of ordered lanes of pedestrians who walk in the same direction and the alternation of streams of pedestrians trying to go through a narrow door in opposite directions.

Other models of pedestrian behavior given by dynamical systems are those related to *mean field games* [215, 189], where a continuum of pedestrians are described as rational agents that maximize some utility function, and move following optimal paths of an underlying optimal control problem. Pedestrians are then characterized by a probability distribution that evolves as a forward Kolmogorov equation, whereas their optimal strategies evolve as a backwards Hamilton-Jacobi-Belmann equation. This forward-backward system is mathematically challenging and has presented a great deal of new insights and results (see *e.g.* [121, 46]).

The models presented here are not exhaustive and we furthermore refer the interested reader to the very complete survey [48], by Castellano *et al.* Furthermore, we avoided on purpose discussing game-theoretical models, such as [313]; since opinion dynamics models and information diffusion models develop theoretical justifications for real-life phenomena such as consensus or clustering formation, and game-theoretical models were introduced in the context of information diffusion, we have chosen to discuss them in detail in the information diffusion introductory section.

0.1.2 Community detection

We observe a variety of different organizations in nature and society: groups of animals, family and friendship networks, protein interaction networks, public transportation networks, etc. The advance of the internet and the mass storage of data allow us to grasp a basic knowledge about these natural or planned organizations; they are nevertheless extremely complex and intricate, hindering our attempts to have a complete understanding of their mechanisms and properties.

One of the first mathematical tools created to infer these networks and study their properties are random graphs [39]. Random graphs are graphs that have a fixed number of nodes (which can converge to infinity as well) and random edges between these nodes. Two major examples of random graphs are the Erdös-Rényi model [96] and the preferential attachment model [21]. In the former, edges are independent and identically distributed Bernoulli random variables (1 if an edge is present, 0 if not) and in the latter they are sequentially attached to random nodes with probabilities proportional to the number of present edges nodes already have. As a consequence, these different behaviors lead to different network properties: for example, in the Erdös-Rényi case one cannot (with a high probability) find nodes with a high concentration of edges, as opposed to the preferential attachment case.

However, real networks are not random graphs, as they display large inhomogeneities, revealing a great deal of order and organization. They present a broad degree distribution and power law tails (see *e.g.* [5]). Furthermore, the edge distribution is not only globally, but also locally inhomogeneous, with high concentrations of edges within special groups of vertices, and low concentrations between these groups. This feature, present in real networks, was pinned *community structure* [116] or clustering.

Hence, among these so-called network properties we have the creation of *communities* or *clusters*. Clustering refers to the phenomenon when nodes of the network can be naturally be grouped into

sets such that each set is densely connected internally, thus dividing the network into smaller groups with dense connections internally and sparser connections externally. As a consequence, pairs of nodes are more likely to be connected if they both belong to the same cluster, and less likely to be connected if not. This clustering phenomenon creates divisions in the network in question, and the understanding of these divisions can lead us to better comprehend not only these objects but the human behavior or the nature itself, as well.

These clusters or communities are important not only for their scientific insights on networked interactions, but also for their concrete real-life applications. For example, clustering web clients having the same interests or being geographically close could help achieve a better performance of internet services [181], identifying groups of purchasers with aligned interests enables to create more efficient recommendation systems [259], grouping proteins having similar functions [258] may shed some light on a better understanding of human and natural cellular mechanisms, etc.

Due to its vast range of applications, there was an explosion of the literature during the last decades, with the presence of complete and thorough tutorials like [102, 267]. They explain a multitude of graph clustering methods, such as spectral methods [234, 235], random walk methods [253, 14], hierarchical clustering algorithms [233, 35], divisive algorithms [116, 236], centrality optimization algorithms [98], methods stemming from statistical mechanics [260], etc. These community detection methods apply tools originating from quite distinct areas of science, such as computer science, statistical physics, social network analysis, spectral analysis, probability theory, optimization theory, and many others.

The community detection techniques developed in the last decades are derived from ancient graph partition methods, such as the max-cut problem [118]. The main difference between older and newer approaches is that algorithms nowadays must be fast and with a low complexity (scaling preferably subquadraticaly in the number of nodes or edges), since a slower but precise algorithm is not even comparable to an approximate but faster one when clustering real-life networks with hundreds of millions of nodes.

Community detection algorithms[1] can thus be grouped into different families, each one highlighting the tools or techniques used to perform such network clustering. We discuss here some of these families: *modularity clustering*, in which authors optimize a global measure of how organized a network is, called modularity [236]; *multiscale clustering*, in which authors adopt methods possessing an adjustable parameter that alters the granularity of the communities found; *random walk clustering*, where authors use probabilistic tools like Markov chains in order to derive clustering algorithms; and *centrality clustering*, where authors develop clustering algorithms based on network centrality measures, such as the edge betweenness centrality [116] or communicability [98] (a centrality measure similar to Katz centrality [168]).

We now study these methods in more details, subdividing each family of community detection methods into smaller subfamilies, starting by modularity clustering techniques.

With the advance of the field of complex networks (large networks without an apparent structure), researchers adopted more statistical qualitative measures of how well clustered a network is, locally and globally. An important measure of clustering is the *modularity* of a network: the modularity compares the original network with a randomized version of itself where edges are randomly rearranged, keeping fixed the expected node degree distribution [236]. Modularity itself is intractable to fully optimize, but there exist less costly alternatives to its complete optimization,

1. We only discuss in this introduction community detection methods for *undirected networks*, i.e., networks such that the edges linking two nodes do not possess a predefined direction and the relationship induced by these edges is symmetric. There exist community detection methods that deal with directed networks (with one-sided relationships between nodes) such as [218], but it is not the focus here.

such as

- *Spectral methods:* the problem of assigning nodes to a fixed number of clusters can be rewritten as a quadratic assignment problem [234, 304], and by applying a convex relaxation being transformed into a quadratic programming problem [234, 304], which allows the use of spectral techniques for this optimization problem. When the assignment is performed for only two clusters, it is basically achieved by calculating the leading eigenvector of the so-called modularity matrix [234]. The assignment of nodes into more than two clusters proceeds in a similar fashion, following the ideas in [234, 235, 304].

- *Hierarchical greedy methods:* where one starts by associating each node with a community and at each step computes the difference in modularity in pairs of communities, continuing in the direction of the higher gain in modularity and merging the associated communities.

  Two famous methods using this hierarchical greedy technique are the method developed by Newman in [233] (which has a similar implementation in [67]) and the Louvain method [35]. The greedy method of Newman [233] associates at first each node to its own community, and at each step it computes the maximum gain in modularity when merging two communities, if any (it may well be that the gain in modularity is negative, so the algorithm terminates); thus, the algorithm merges the two communities with the maximum gain in modularity, and proceeds in an agglomerative fashion until termination at a local maximum of the modularity, creating a network dendrogram.

  The Louvain method, developed by Blondel *et al.* in [35], works in a similar fashion to Newman's greedy method. It is a greedy optimization of the modularity, using smartly the fact that its local computation is quite fast. It starts, again, by associating each node with its own community, and consists on the repeated iteration of two steps: first, it sequentially sweeps over the nodes and given a node, it computes the difference in modularity of inserting it in a neighbor's community, performing the insertion of nodes in the community with the higher increase in modularity, if any. In a second step, the nodes belonging to the same community are merged, with the new community edge weights the sum of the weights of the edges of the underlying nodes. These steps are repeated, generating a dendrogram from which one can choose the best community structure. This method is extremely fast and yields higher global modularity than other greedy techniques, but it has been remarked that it is still not clear whether some of the intermediate parts of the associated dendrogram are indeed meaningful hierarchical levels of the graph [102].

Despite its great success, the modularity optimization approach suffers from a *resolution limit* problem, where it may fail to find communities smaller than a given scale which depends principally on the network's size, as pointed out by Fortunato and Barthélemy [103]. The same remark was made by Kumpula *et al.* [183] regarding other null models.

In order to overcome this resolution limit problem, multiscale (or multiresolution) methods were introduced. These methods possess an adjustable parameter that helps tuning the granularity level of the communities found. They can be divided (for example) into

- *Modularity optimization methods:* Arenas *et al.* discuss in [11] this resolution limit not as a problem, but as an intrinsic property of the network, proposing a multiscale method by introducing self-loops with different weights in the original network and performing again a modularity optimization approach.

- *Statistical physics methods:* Reichardt and Bornholdt have shown in [260] (with the weighted case being presented in [143]) that it is possible to reformulate the problem of community detection as a problem of finding the fundamental state of a Potts model [307], where each node is described with a spin. Thus, their method consists in minimizing a Hamiltonian function that takes into consideration, as in the modularity case, a difference between a null model with the same degree distribution of the original network and the actual interaction between nodes, with a scale parameter $\gamma > 0$ responsible for the granularity of the communities found. If $\gamma = 1$, they recover (up to a multiplicative factor) the modularity function of [236].

- *Signal-processing methods:* Different community detection algorithms appear when signal processing tools are used, such as wavelets [217]. These techniques are successfully used in many fields, such as the detection of subgraphs in networks [226].

  A method that exploits the fact that networks may have different community structures when using different resolutions is the wavelets-based community detection algorithm developed by Tremblay and Borgnat in [288], where the authors use band-pass filters defined in the graph Fourier domain, generated by stretching a band-pass filter kernel with a scale parameter $s > 0$. They thus use the Fourier modes of the graph Laplacian (its eigenvectors) to create the wavelet base (for a fixed resolution parameter $s > 0$) and draw a dendrogram of the target graph by the node correlation between the elements of the base. The higher the resolution parameter s, i.e., at larger scales, the fuzzier the resolution, i.e., we have bigger communities; the lower the resolution parameter s, the wavelets use higher frequency modes and therefore create a higher localization, generating smaller communities.

Another community detection framework is one based on random walks on networks [213]. Let us assume, without loss of generality, that each node has at least one neighbor, and let us define the probability of a random walker going from one node to another to be proportional to the weight of the edge linking these two nodes, i.e., if a random walker is placed in a given node, the probability of him going from this node to a neighbor node is proportional to the edge weight. If there is no edge linking a pair of nodes, the random walker cannot go from one node to the other (at least in only one hop). Classical Markov chain techniques allow the study of this random walk in depth, such as discovering the analytic form of its stationary distribution (under certain hypothesis) and the mixing time, and are closely related to spectral properties of the Laplacian matrix of the underlying network [213].

Hence, one may expect that random walks may indeed help in discovering divisions and nontrivial structures in networks. For instance, Van Dongen stated in his Ph. D. thesis [293] some basic principles behind random-walk network clustering algorithms: 1) The number of higher-length paths in a network is large for pairs of vertices lying in the same dense cluster, and small for pairs of vertices belonging to different clusters. 2) A random walker that visits a dense cluster will likely not leave the cluster until many of its vertices have been visited. 3) Considering all shortest paths between all pairs of vertices, links between different dense clusters are likely to be in many shortest paths.

We now discuss in detail some community detection methods that make use of random walks in networks to uncover cluster structures:

- The first algorithm to be presented here is the Markov clustering algorithm (MCL) of Van Dongen [293]. The algorithm consists in the iteration of two steps: a first step called expansion, in which the random walk transition matrix is raised to an integer power p (it is well known that the resulting transition matrix gives the probability that the random walker goes from

one node to another in p steps), and a second step, called inflation, consisting in raising each entry of the new transition matrix to some real-valued power $\alpha > 0$ and renormalizing the new matrix to be again a transition matrix of a random walk. This last step enhances the weights between pairs of vertices with large values, which are likely to belong to the same community. After a few iterations, the process usually delivers a stable matrix, with some remarkable properties. Its elements are either zero or one, and the generated network described by this matrix is disconnected, with its connected components the uncovered communities of the original graph.

- A particularly interesting instance of the MCL is the label propagation algorithm introduced by Raghavan *et al.* in [256]. In the MCL, the author mentions a possible reduction by keeping only the k maximum nonzero entries of each column after the expansion steps, which is taken to the extreme by the label propagation algorithm: one keeps only the largest entry of each row after each expansion step. The resulting algorithm can be designed as follows: one starts with each node belonging to its own community; at each step, in a randomized order, each node is assigned to the community containing the higher number of its neighbors (if there are more than one such communities, one decides randomly among them); the process is repeated until no more assignments can be done (it is worth mentioning that the algorithm may not end).

  As a consequence, the resulting communities have nodes that possess more edges with vertices inside its community than vertices with nodes in each of the other communities, compared in a pairwise fashion. Tibély and Kertész prove in [283] that the label propagation algorithm is equivalent to finding the fundamental state of a zero-temperature Potts model [307], giving a precise description of the communities found.

- Pons and Latapy show in [253] that the entrapment of a random walker in clusters can be measured by quantities related to spectral properties of the transition matrix, defining distances between vertices and communities. The resulting algorithm, called *Walktrap*, proceeds in the following greedy fashion: it starts by assigning each node to its own community (as usual), and at each step it calculates the distance between (every reasonable) two communities and merges the two communities with the smallest distance, following Ward's method [301]. It provides then a hierarchical structure of the target network.

- Finally, Avrachenkov *et al.* develop in [14] a community detection algorithm based on the mixing time of local random walks in a graph. Their algorithm computes for each cluster a scoring function that takes into consideration the spectral gap of the transition matrix of a random walker moving only inside the cluster in question (which is a proxy of how fast the random walk mixes, i.e., converges to the stationary distribution, and measures how well connected the cluster is) and the probability of not leaving the cluster if started inside it (which is a proxy of how disconnected the cluster is with the rest of the network). The algorithm thus performs an aggregating search for communities starting at each node being its own community. This leads to a dendrogram representing the community structure of the graph in question.

We conclude the discussion of community detection methods with centrality-based methods, which find structures in networks using centrality measures. Some of these methods are the following:

- The betweenness centrality method develop by Girvan and Newman in [116], which extends the concept of betweenness[2] from vertices to edges, and leverages this idea to create a divisive algorithm based on the idea that edges linking different communities usually have high betweenness as paths from different communities must pass through them.

  The algorithm begins with the whole network as a unique community, and at each iteration it removes the edge with the highest betweenness, recalculating the new edge betweenness until no edges remain and uncovering communities that eventually separate from one another. This algorithm, as one might notice, constructs a dendrogram of the network where one may cut it in any desirable fashion in order to retrieve the communities.

- An entirely unrelated approach, which resembles random-walk-based techniques, is given by Ernesto and Hatano in [98], which is based on the generalization of the *communicability* of nodes taking into account not only shortest paths between nodes but also the other paths linking them.

  The authors analyze a matrix taking into consideration all paths linking nodes in an unweighted network (powers of the original adjacency matrix), studying the Green function given by the weighted sum of the number of paths. Thus, using spectral properties of this matrix the authors are able to define rules to assign each node to communities.

The discussion led in this introductory section is not exhaustive and we refer the interested reader to the survey in [102] where Fortunato presents in detail methods for undirected networks; [218], where Malliaros and Vazirgiannis illustrate the theory and methods for directed networks; and [252], where Platié and Crampes present a survey featuring a new approach based on semantics, which allows interpreting social relations in community detection for social networks.

0.2 Information diffusion and trend detection

The second part of this thesis is dedicated to an *information diffusion* model and a subsequent *trend detection* algorithm using ideas from one of its particular cases. We provide now a general review of the literature about both subjects.

0.2.1 Information diffusion

Information diffusion models study the broadcasting and adoption of information by users in networks. The basic idea is that people want to talk, share, post, tweet, like and perform any other kind of social action in social networks, and by doing so, they influence others to do the same. For example, when some new technology arrives, early buyers post photos and comments on social networks, which are then diffused by "word-of-mouth". These actions lead, in turn, to an increase in the number of buyers for this new technology, and so on. In Twitter, when someone retweets a piece of news, he first must have been in contact with the news itself, so the broadcasting of the initial news creates a cascade of new notifications and posts; these objects are called *information cascades* [93].

The first information cascade models were opinion-dynamic and game-theoretical based models, in which agents in a network have information they want to spread, or a utility function they want

2. For a given node, its vertex betweenness [104] is defined as the sum, among all pairs of other nodes, of the ratio between the number of shortest paths (for unweighted networks) containing the given node and the total number of shortest paths.

to maximize. These models were quite useful when studying political trends [268], cultural fads and fashion changes [26], collective actions [127], viral marketing [263], diffusion of innovations [264], etc.

The division of these information diffusion models in opinion-dynamic versus game-theoretical models is due basically to the difference of how agents interact: in opinion-dynamic based models, agents do not seek to maximize a utility function; these models have simple rules and heuristics for agents' interactions and the goal is to study the aggregate behavior of the system. Some examples of opinion-dynamic based model are *independent cascade models* [119, 120], *threshold models* [302, 24], *interacting particle systems* [203, 7] (which are the continuous-time analogue of the stochastic cellular automaton [295]), and other frameworks already mentioned during the general introduction about opinion dynamic models.

Let us now present them in more detail:

- *Independent cascade models:* Are stochastic models of information diffusion, in which there exists an initial seed group possessing some information that propagates through the network from agents to their neighbors, until no agent is able to continue the transmission process [119, 120].

  Agents can be in two different states, active or inactive. An active agent may at each time step disseminate the information to a random inactive neighbor, such that the transmission occurs with a probability depending on their tie. The information transmission has only one chance to succeed, i.e., if an active agent fails to transmit the information to an inactive neighbor, it cannot repeat it to this same neighbor. An inactive agent can become active if it successfully receives the information from an active node, and upon being activated it remains active until the end of the diffusion process.

  Thus, at each discrete time step, randomly chosen active agents try to disseminate their information to some of their inactive neighbors, in an independent fashion, i.e., the transmissions are independent from each other.

  Independent cascade models are in the core of some influence maximization problems, such as [169].

- *Threshold models:* These models are also based on stochastic diffusion of information, but in the opposite sense of cascade models. There are also two states for each agent, active and inactive, and the model begins with an initial seed group disseminating some information. The difference from Independent cascade models happens in the sense of transmission, i.e., at each time step the inactive agents compare the number of active neighbors against the number of inactive ones, and become active depending on whether it is greater than some predefined influence/infection threshold (which may be agent-specific [127] or fixed for the entire network [24]). Usually this rule takes into consideration the weights on the edges of the underlying graph, i.e. the bigger the weight in an edge, the larger the influence of a neighbor.

  The classification of the subsequent model depends on how this influence/infection threshold is defined. For example, majority threshold models [247] dictate that the information is successfully transmitted to a given inactive agent if the (weighted) majority of his neighbors are active; linear threshold models [216, 302] are themselves based on the assumption that for a given inactive agent the (weighted) proportion of his active neighbors must be bigger than an agent-specific threshold in order for the transmission to take place; and fixed-value threshold models [24] follow the same ideas of linear threshold models, but with a fixed

infection threshold for all agents in the network (for a fixed threshold of 1/2 it becomes the majority rule).

Although opinion-dynamics-based model are well studied and possess a large literature in information diffusion, they are still based on predefined rules and simple actions of agents, which have a different interpretation in economics. The premise is that agents happen to adopt these transmission rules and dynamics because they are rational and seek to maximize some sort of utility or gain, which comes in different forms. Some examples of game-theoretical information diffusion models are:

- *Stochastic best-response dynamics:* The assumption in these models is that agents are not fully rational when making a decision (the bounded rationality hypothesis [224]), such that the actions are played following certain probabilities [37, 95].

  At each (random) time agents revise their strategies and need to select an action to play, which is taken to be the best-response of a game composed by their neighbors and the actions chosen by them. As agents are not fully rational, they are not entirely certain that this game reflects their true utility and thus, select a random action following a probability distribution depending on the future utility of each action.

  When this probability function is the softmax function [28], this mechanism is called *log-linear dynamics*, and has been extensively studied in the game theory literature [225, 9].

- *Diffusion of innovations:* In this particular model the dynamics are similar to that of stochastic best-response models, in the sense that agents are not fully rational and assign probabilities to actions, which again depend on the future utilities stemming from local interactions with their neighbors [313, 229].

  The major difference between the two models is the fact that the choice of agents here is not dependent on the present actions of neighbors (thus not playing a classical game) but on their last played actions. As stochastic best-response dynamics, when the choice probability function is the softmax distribution, it has been shown that under mild assumptions on the utility functions and the network properties the agents' actions configuration converges to a stationary state [37]. When players become increasingly rational (when the noise parameter of the softmax distribution converges to zero) the stationary state gives rise to the so-called *stochastically stable states* [312], which are agents' pure actions that are played with a nonzero probability when agents are fully rational, i.e., they maximize agents utilities.

- *Network games:* In a more general fashion, the game-theoretical mechanism of agents playing games in networks has been coined *network games*, which could have different forms: diffusion of behavior and information cascades [111, 194], network formation games [159], etc. The reader can be referred to [160] for an extensive review of literature on the subject.

Apart from the beforementioned theoretical models such as opinion dynamics and game-theoretical models for information diffusion, a fruitful new research program came along in the past years, such as Kempe *et al.* in [169], where the authors study how to maximize the initial seed set able to create the largest information cascade in a social network, assuming that nodes pass along the received information as in an independent cascade model or a linear threshold model. After the pioneering work of Kempe *et al.*, new and more complex information diffusion models started to be developed, with a greater emphasis on the algorithmic part.

These models take into consideration different aspects of information diffusion: the diffusion patterns and times [317, 99], the contents diffused and their dissemination probabilities [195, 231], the role of users at diffusing these contents [125, 201], the temporal shape of the impact/influence of these diffusions [195, 232], the reconstruction of the networks given the observed cascades [123, 126, 316], and many other properties of this complex process.

Instead of modeling the qualitative properties of the dissemination process itself, these works focused on retrieving the network properties from the likelihood of the information cascades. Their goal is hence twofold: first, by retrieving the system parameters, these models are able to obtain crucial information on users and the information being disseminated; second, by choosing a parametric model of information diffusion, they still model the diffusion process itself from the estimated parameters. They thus provide a more complete approach to the information diffusion process by not only modeling it but at the same time retrieving vital information about the network and the disseminated contents.

Although a complete division of these works is rather difficult due to their diversity, we can divide them into two categories: those that do not use point processes [74] and those that use them. The main difference between them is that those that do not use point processes base themselves on simple heuristics and empirical properties of information cascades to derive exploratory models. For example:

- Leskovec *et al.* report in [197] some findings about the linking of blogs and the structure of the information cascades, after analyzing a dataset with $45,000$ blogs and $2,2$ million blog postings, designing a simple flu-like epidemiological model that mimics the spread of information and produces information cascades similar to real-life ones.

- Leskovec *et al.* develop in [195] a scalable framework for meme-tracking[3], providing a representation of the news cycle. They identify a large class of memes exhibiting wide variation on a daily basis by tracking $1,6$ million mainstream media sites and blogs over a period of three months with a total of 90 million articles, finding new and persistent temporal patterns for the information diffusion in such contexts.

- Myers and Leskovec study in [231] the variation of the probability in retransmiting information due to previous exposure to different types of information; they found that, for Twitter, these retransmission probabilities are indeed very different when compared to results stemming from independent cascade models, which reinforces the discussion in deriving new model-free approaches to information diffusion with multiple contents.

- Myers *et al.* study in [232] the influence of externalities over nodes on information cascades in networks, adopting a cascade model with parameters relating to dissemination of information from external sources and internal sources of social networks, for which the time instances of diffusion are essential to the maximum likelihood estimation procedure.

- Snowsill *et al.* develop in [273] a network inference algorithm based on text mining techniques in order to associate markers with reused text, in order to track pieces of text that travel through nodes in a network. The reconstruction of the infection network is thus performed using an approximated set covering technique [60] developed in [107] to infer a minimal graph spanning a suitable stochastic branching process.

3. A *meme* is an idea, behavior, or style that spreads from person to person within a culture [79].

- Gomez-Rodriguez *et al.* [122] and Daneshmand *et al.* [76] propose diffusion network inference algorithms for recovering the diffusion probabilities in continuous-time cascade models, and Gomez-Rodriguez *et al.* develop in [123] an algorithm capable of uncovering the most-likely network configuration leading to a given information cascade, following the independent cascade model already presented in this introduction.

On the other hand, the works that use point process techniques base themselves on parametric point processes in order to "guess" the interactions between users and information, estimating their parametric version of the reality from the likelihood of events. Differently from other models, point-process-based modeling has the advantage of aggregating several properties of not only the information diffusion process itself also but the network, thus working as a trade-off between a simpler view of reality against a more complete overview of the diffusion process. For example, a point process that is deeply studied in the second part of this thesis is the Hawkes process [140, 208], which is a self-exciting point process possessing a parametric intensity that takes into account the previous events to increase the likelihood of future ones.

Some of its examples are:

- [38], where Blundell *et al.* model reciprocating relationships with Hawkes processes [140, 208], using a Bayesian nonparametric model that discovers the implicit social structure from interacting data, based on the Infinite Relational Model [309].

- In [158], Iwata *et al.* propose a probabilistic model for discovering users latent influence using cascades of inhomogeneous Poisson processes. The authors present a Bayesian inference procedure of the model based on a stochastic expectation-maximization algorithm.

- Gomez-Rodriguez *et al.* generalize in [124] independent cascade models with survival theory, developing general additive and multiplicative diffusion models. The proposed framework solves efficiently the inference of the most probable network responsible for such cascades.

- Yang and Zha study in [310] the propagation of memes in social networks with linear Hawkes processes and couple the point process with a language model in order to estimate the memes. They provide a variational Bayes algorithm for the coupled estimation of the language model, the influence of users and their intrinsic diffusion rates.

- Zhou *et al.* develop in [317, 316] a model for the information diffusion process with a multivariate Hawkes process, developing parametric and nonparametric learning algorithms for the system parameters from the cascades of data.

- Li and Zha develop in [202] a compact learning algorithm for Hawkes-based parametric models of information diffusion, replacing the influence kernels between users and information with sparse temporal linear combinations of a reduced number of features. The authors compared the proposed compact method with classical Hawkes learning methods in order to assess the gain in predictive power when dealing with an insufficient number of information cascades.

As one can notice, this new wave of literature on information diffusion is quite extensive, with new methods being developed using different sets of techniques such as cascade models, probability theory, statistics, point processes, language models, etc.

We refer the interested reader to a well explained tutorial by Guille *et al.* [132] for more details.

0.2.2 Trend detection

Finally, we present a general overview of the literature related to the last chapter of this thesis, *trend detection*. Trend detection is the study of models, techniques and tools used for detecting or predicting patterns in information. Trends come in various forms and shapes; for example, one may look at the return time series of a stock and try to denoise it, revealing a deterministic trend [101]; one may be concerned about anomaly detection [4], where one looks at outlier events in some data, forecasting which events are defined as "normal" and which ones are defined as "exceptional", etc.

The literature on topic/trend detection is very heterogeneous (see *e.g.* [8] for an overview), employing several methods such as information diffusion methods [169, 58], adoption models and complex contagion models [52, 51], text mining and queuing techniques [173, 300], etc.

The methods using information diffusion in social networks are pioneered by Kempe *et al.* in their seminal paper [169], where the authors develop an optimization framework to study the problem of identifying influential users for a marketing campaign, introduced by Domingos and Richardson in [88]. The framework uses submodular functions to detect the optimal seed group in order to diffuse a content, based on the already explained independent cascade propagation model [119, 120] and linear threshold model [216, 302].

The problem of finding the best seed group that maximizes the expected number of influenced nodes was denoted *influence maximization problem*, and inspired numerous works:

- S. Bharathi *et al.* [25], where the authors extend the influence maximization problem for multiple competing topics using game-theoretical techniques, and N. Barbieri *et al.* [22], where the authors extend the independent cascade and linear threshold models to take into account multiple topics, and devise a new influence propagation model that instead of considering user-user influences, leverages user authoritativeness and users interests in topics, leading to a compact representation of parameters.

- Chen *et al.* derive in [56, 57] scalable extensions of the independent cascade and threshold models, and Tang *et al.* derive in [279] an influence maximization algorithm that has near-optimal complexity based on the triggering model defined in [169], which is an extension of the classical independent cascade and linear threshold models.

- Chen *et al.* develop in [55] an influence maximization algorithm based on an extension of the independent cascade model taking into account not only positive opinions, but also negative ones.

- Zhuang *et al.* study in [318] the influence maximization problem on dynamic social networks. Their setting consists in making periodical partial observations of the social network, with the derivation of an algorithm that minimizes the difference between the expected number of influenced nodes in the real social network and in the partially observed one, under the classical independent cascade model.

- Gomez-Rodriguez and Schölkopf developed in [125] a greedy influence maximization algorithm taking into consideration the continuous-time independent cascade model of [122].

Although the influence maximization literature is quite extensive and very important for the understanding of trend detection in social networks, it does not represent its totality. For example, some works using text mining and queuing theory are

- [173], where Kleinberg develops a trend detection algorithm by modeling "bursts of activity" over document streams using an infinite-state automaton (analogous to models in queuing theory for bursty network traffic), in which bursts appear naturally as state transitions.

- [300], where Wang *et al.* propose a general probabilistic algorithm that discovers correlated bursty patterns and their periods across text streams, even if the streams have completely different vocabularies - *e.g.* English vs. Chinese.

- [10], where AlSumait *et al.* propose an online topic model based on the Latent Dirichlet Allocation [34], a generative hierarchical Bayesian model for text data, serving as foundation to an algorithm detecting bursty topics in social networks. The idea is to incrementally update the topic model at each time step by using the previously generated model, creating thus a temporal evolutionary matrix for each topic and permitting the detection of bursty topics.

With the advent of social networks and their tsunami of data, a vast corpus of empirical works appeared in the literature. They not only shed light into sociological phenomena [303], but also on the information diffusion process [52, 51] and appearance of trends in social networks [306], acting as a bridge and decreasing the gap between the abstract models of academia and real-life stylized facts of social networks.

In spite of the fact that these empirical works do not reflect directly trends and do not develop trend detection algorithms, we have chosen to discuss them in this section because they represent the interaction between the theoretical works on the information diffusion models used to devise better trend detection algorithms and the real-life data to support such theories. They provide thus precious insights that allow the manufacturing of new and more realistic information diffusion models, and as a consequence, better trend detection algorithms. Some of these empirical works are:

- [157], where Huberman and Adamic discuss several studies of information flow in social networks. They uncover critical properties of social networks and the information diffusion process, such as their underlying social structure, how information spreads and why small world experiments give solid results.

- [3], in which Adar and Adamic create a tracking algorithm responsible for discovering the information flow in the blogshpere using several features of pairs of blogs: the number of common blogs explicitly linked to by both blogs, the number of non-blog links shared by both of them, text similarity, order and frequency of repeated infections, and in-link and out-link counts for both of them. They also created a visualization tool in order to get a better understanding of the diffusion process.

- Wu and Huberman study in [306] how attention to novel items propagates and fades among large populations. They analyzed the dynamics of 1 million users of the social network Digg and described it by a temporal model with single novelty factor.

- Centola *et al.* empirically illustrate in [52, 51] that a complex contagion model is more precise than simple adoption models [86] for information diffusion in social networks, studying the qualitative effects of network topology on its ability to propagate collective behavior.

- Gao *et al.* study in [112] real anomalous events using mobile phone data, and find that information flow during emergencies is dominated by repeated communications.

- Bakshy *et al.* [20], Teng *et al.* [282] and Weng *et al.* [303] analyze how the social networks can influence the diffusion of different topics, and vice-versa, i.e., given different kinds of topics, the relevance and creation of strong and weak ties in the social networks during the diffusion process.

All the works presented so far tackle the trend detection problem at some instance, either in a theoretical or practical manner. However, they are not perfectly suited to deal with trend detection in social networks, as they do not exploit most of the relationships between users and information, as for example *user social authority and influence*, *topic influences*, *information flows*, *social actions*, *contextual data*, etc. We now present some of the works dealing with these issues:

- Cataldi *et al.* devise in [50] an algorithm to detect in real-time emerging topics on Twitter. First, they extract the contents of the tweets with a model for their life cycle. Second, they consider the social importance of the sources of the tweets, using the Page Rank algorithm to analyze the social ties of users. And finally, they create a topic graph connecting the emerging terms with other semantically related keywords.

- Takahashi *et al.* derive in [276] a trend detection algorithm focusing on the social aspects of social networks, with links between users being generated dynamically through replies, mentions, etc. The authors propose a stochastic model for behavior of a social network user, detecting the emergence of a new topic. They combine the proposed anomaly score with a change-point detection technique based on the Sequentially Discounting Normalized Maximum Likelihood coding [291], or with Kleinberg's burst model [173].

- Budak *et al.* define and identify in [45] coordinated trends (characterized by the number of connected users discussing them) and uncoordinated trends (characterized by the number of unrelated people interested in them), providing network-oriented solutions for detection of such trends.

- Guille and Hacid derive in [131] an algorithm based on the asynchronous independent cascade model [266], using an information diffusion model that captures and predicts the dissemination process, relying on semantic, social, and time features.

- Guille and Favre devise in [130] a Mention-Anomaly-Based Event Detection algorithm on Twitter, based on the creation frequency of dynamic links users insert in tweets to detect important events and estimate their magnitude. The proposed algorithm dynamically estimates the time periods for the events, not assuming them of fixed duration.

- Cheng *et al.* propose in [58] a framework for addressing cascade prediction problems, motivated by a view of cascades as complex dynamic objects passing through successive stages while growing.

Part I

Opinion Dynamics and Community Detection

CHAPTER 1

Opinion dynamics

"Opinions are made to be changed - or how is truth to be got at?"

— Lord Byron

1.1 Introduction

We begin this thesis with a theoretical model of information diffusion in social networks, based on an opinion dynamics model. This model is necessary to lay the fundamentals of information diffusion ideas in order to fully explore them during the course of this thesis. Thus, the first part of this thesis will have a theoretical and abstract flavor at first, and will be followed by an application of this theoretical model: an opinion-dynamics-based community detection algorithm.

Opinion dynamics models develop rules on how a group of agents communicate and analyze their impact at the network level (see [48] for a survey): do these rules lead to network consensus, spontaneous clustering, etc.? Interestingly enough, simple opinion dynamics models often suffice to be confronted with deep technical issues (see *e.g.* [53]) and fascinating conjectures (see *e.g.* [142]); in addition, they also cover a large number of real-life situations that possess nontrivial behavior: flocks of birds [72], interacting groups of people [105], distributed systems of robots [85], physical particles with spins [254], etc.

This chapter introduces an opinion dynamics model based on exchange of opinions between agents over multiple contents, and studies its convergence. We model the system as a weighted, undirected graph, consisting of N agents. All agents have a common set of contents, say K contents, and each agent maintains a vector of scores, each of which reflecting the instantaneous appreciation of the agent for each content. This appreciation starts with the agent's own initial opinion for each content, and then evolves as a function of signals the agent receives from her neighbors. Each signal consists of the identity of one content, a number from 1 to K, that the neighbor chooses to broadcast to the agent; the agent then updates its score for the specific content by the weight of the link she has to the broadcasting neighbor. This choice is done by a random sampling from the agents normalized scores on the contents, using a nonlinear transformation - the softmax probability function with parameter β [28].

The softmax parameter β impacts the choice of the signal: a small value of β corresponds to a uniform choice over the K contents, and as β grows, the sampling becomes more biased towards contents with larger scores. We show that for each fixed and finite value of β, we obtain convergence for the agents normalized scores as time tends to infinity.

The sociological broadcasting/sharing mechanism used to construct our model has its roots in the seminal work of DeGroot [83], modeling agents in a network possessing a scalar opinion about a subject, interacting with neighbors through combinations of opinions coming from themselves and from their neighbors. This same broadcasting idea was further developed by Tsitsiklis [289], Boyd *et al.* [43], and many other authors to create decentralized gossip schemes, which aim to compute linear functions of agents scalar opinions/states through pairwise interactions of agents, resulting in the convergence of opinions towards a consensus. However, the convergence in our model, unlike in standard consensus algorithms, may lead to disagreement between agents over the contents: since a large softmax parameter β generates a greater bias towards contents with larger scores during the agents' random sampling, agents tend to broadcast opinions about contents for which they have a greater appreciation; thus, depending on the way agents are connected, each agent starts receiving more often signals about contents she appreciates the most, hence partaking a larger appreciation for this same content with a subgroup of her neighbors, *i.e.*, the network ends up clustered over contents most appreciated by agents.

A similar clustering phenomenon also appears in a class of nonlinear opinion dynamics models, the so-called bounded confidence models [142, 82], where agents are placed in a communication network and possess two quantities, the first one being a scalar opinion/state and the second one being a confidence interval, which is a time-changing symmetric interval centered around the instantaneous value of her opinion. Each agent's opinion evolves thus through interactions with the neighbors that reside inside the agent's confidence interval, *i.e.*, agents only interact when they present opinions sufficiently close to each other's. Since, at each time step, agents only interact with neighbors possessing an opinion sufficiently similar to their own, agents' opinions polarize and converge towards different clusters, as expected. Although our model presents the possibility of clustering, as in bounded confidence models, the interaction between agents are fundamentally different in both cases: in bounded confidence models agents update their opinions following rules similar to gossip schemes, using linear combinations of their actual opinions and the opinions of neighbors residing in their confidence intervals, which results in similar opinions becoming even more similar, while in our work agents do not have any control over the information received from neighbors and the eventual clustering is an indirect effect, consequence of neighbors finally being more appreciative, in the long run, of a common content.

As our model is based on scores measuring the appreciation of contents by agents, a reinforcement mechanism takes place: the more agents broadcast opinions relative to a content, the higher the probability their neighbors have of broadcasting opinions about the same content back to them. This behavior is not "hard" or "binary" as in classical voter models [68, 205], in which agents adopt different states, but "soft" in the sense that the changes in agents' opinions happen in a gradual fashion. Due to the smooth changes in agents' opinions, our model resembles recent voter models with reinforcement [75, 49, 178]. Nevertheless, in our model, this reinforcement happens "on the contrary sense" of the literature, as in the Sznajd model [275], which has an "outflow" dynamics where each agent propagates her binary state to one of her two neighbors when she possesses the same state of her other neighbor. This "outflow" dynamics appears as well in the proposed model since agents broadcast their opinions to neighbors, with the reinforcement occurring in an indirect level, when agents receive broadcasts from neighbors that randomly select the broadcasting information from a softmax probability function.

Our model

i) takes into account the presence of multiple contents,

ii) relies on the random sampling of the broadcasted contents performed by agents through the

nonlinear softmax function with parameter β, and

iii) uses a stochastic broadcasting scheme of information where each agent transmits at each time step an opinion about a single content to her neighbors.

Moreover, we prove the convergence of the normalized scores of agents, for any finite and fixed value of β, towards a particular set, however it turns out that studying this set is a challenging task. For example, when the value of the softmax parameter β is small, agents' normalized scores converge to a consensus, whereas when β is large, one can observe through numerical examples that the convergence of agents' normalized scores lead to the clustering of the network into groups where agents have the same preferred content.

The rest of the chapter is organized as follows. Section 1.2 describes the proposed opinion dynamics model and our main convergence result. In Section 1.3 we provide a mathematical proof for the convergence of our opinion dynamics algorithm. In Section 1.4 numerical experiments are performed to support our claims and Section 1.5 concludes the chapter.

1.2 Model description and main result

1.2.1 Notations

For two rectangular real matrices $P, Q \in \mathcal{M}_{N\times K}(\mathbb{R})$, let $\langle P, Q\rangle = Tr(P^T Q)$ be their scalar product, with associated norm $||P|| = \sqrt{\langle P, P\rangle}$. Also, let 1 be the vector with entries 1, where the dimension of the vector is clear from the context, and let us define for a matrix $P \in \mathcal{M}_{N\times K}(\mathbb{R})$ the vectors $P^i = (P^{i,1}, \cdots, P^{i,K}) \in \mathbb{R}^K$.

We say that a sequence $(m_t)_{t\in\mathbb{N}} \in \mathcal{M}_{N\times K}(\mathbb{R})$ converges to the set $\mathcal{E} \subset \mathcal{M}_{N\times K}(\mathbb{R})$ if and only if $d(m_t, \mathcal{E}) \to 0$ when $t \to \infty$, where $d(m_t, \mathcal{E}) = \inf_{z\in\mathcal{E}} ||m_t - z||$ is the distance between m_t and the set $\mathcal{E}$.

Let us define the $(K-1)$-dimensional simplex $\Delta_K = \{x \in \mathbb{R}_+^K \mid \sum_{k=1}^K x^k = 1\}$ and Δ_K^N the set of $N \times K$ real matrices such that every row is in Δ_K (the set of $N \times K$ real stochastic matrices), i.e.: $\Delta_K^N = \{M \in \mathcal{M}_{N\times K}(\mathbb{R}_+) \mid \sum_{k=1}^K M^{i,k} = 1,\ \forall i \leq N\}$. Also define the set $\mathring{\Delta}_K^N = \{x \in \mathcal{M}_{N\times K}(\mathbb{R}) \mid x1 = 1 \text{ and } x_{i,k} > 0,\ \forall (i,k)\}$ and, for a parameter $\beta > 0$, the entropy function $\mathcal{H}_\beta : \Delta_K \to \mathbb{R}_-$ as $\mathcal{H}_\beta(y) = \frac{1}{\beta}\sum_k y^k \log y^k$.

1.2.2 The opinion dynamics model

We begin by presenting in detail our opinion dynamics model: let us consider a network of N agents sharing opinions about K distinct contents with their neighbors. The network could consist of a small group of agents chatting in the same room or a large social network [180, 2, 78], and these contents might be an ensemble of movies, books, political leaders, etc. The communication network is coded by a weighted, undirected communication graph $G = (V, E)$ that represents the network topology, where $V = \{1, \ldots, N\}$ is the set of agents and E is a subset of all possible communication links between agents, such that the link $i \sim j$ belongs to E if and only if agents i and j are able to communicate with each other.

The weights of G represent the influences that agents have over one another: if agent j is very influential over agent i, then the broadcasts of agent j have considerable impact over agent i, which is hence reflected by a large weight for the link $i \sim j$. Mathematically speaking, the weights of G are the entries of the $N \times N$ symmetric matrix A, the adjacency matrix of V such that $A_{ij} > 0$ if

and only if $i \sim j$. Defining the $N \times N$ diagonal degree matrix D such that $D_{ii} = \sum_j A_{ij}$, we have that the entry D_{ii} represents the total influence of the network over agent $i \in V$.

As already explained, our model assumes that every broadcast is relative to one of K distinct contents, each of them denoted by an integer $k \in \{1, 2, \cdots, K\}$, with each agent i possessing a specific score $X_t^{i,k} \in \mathbb{R}_+$ for content k at time $t \in \mathbb{N}$. The higher the score $X_t^{i,k}$, the higher is the appreciation for content k by agent i at time t, which in turn implies that agent i is more likely to perform a broadcast relative to content k to her peers. More specifically, we can define $X_t \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$ as the matrix of scores of all agents, such that $X_t^{i,k}$ is the score of agent i over content k at time t, and we can define $P_t \in \Delta_K^N$ as the agents normalized scores, where

$$P_t^{i,k} = \frac{X_t^{i,k}}{\sum_{k'} X_t^{i,k'}}.$$

In our model, agent i selects the content relative to her broadcast at time $t+1$ according to a random variable $I_{t+1}^i \in \{1, \cdots, K\}$ following the random law

$$\mathbb{P}(I_{t+1}^i = k | \mathcal{G}_t) = f_\beta^{i,k}(P_t) \tag{1.1}$$

where

- $P_t \in \Delta_K^N$ is the matrix of agents' normalized scores, defined as $X_t = D(X_t)P_t$, with $D(X_t)$ the diagonal degree matrix of scores X_t defined by $D(X_t)_{ii} = \sum_{k'} X_t^{i,k'}$,
- $f_\beta : \mathcal{M}_{N\times K}(\mathbb{R}) \to \Delta_K^N$ is the softmax function [28] with parameter β, defined as

$$f_\beta^{i,k}(p) = \frac{e^{\beta p^{i,k}}}{\sum_{k'} e^{\beta p^{i,k'}}},$$

  where $p \in \mathcal{M}_{N\times K}(\mathbb{R})$ is a $N \times K$ matrix and $p^{i,k}$ is its (i,k) entry, and
- $\mathcal{G}_t = \sigma(X_s,\ s \le t)$ is the standard filtration associated with X.

Then, after the broadcasting phase at time $t+1$ where agents perform the random sampling and broadcast the selected information to neighbors, agents interpret the received information and update their own scores X. A typical agent i groups every information broadcasted to her and updates her scores accordingly: she adds A_{ij} to the score relative to the content brodcasted by her neighbor j, i.e., to the score relative to the content given by the random variable $I_{t+1}^j \in \{1, 2, \cdots, K\}$.

We have thus the following update mechanism for the agents' scores at time $t+1$:

$$\begin{aligned} X_{t+1}^{i,k} &= X_t^{i,k} + \sum_{j\sim i} A_{ij} \mathbb{I}_{\{I_{t+1}^j = k\}} \\ &= X_t^{i,k} + \sum_{j} A_{ij} \mathbb{I}_{\{I_{t+1}^j = k\}}, \end{aligned} \tag{1.2}$$

which can be written in matrix form as

$$X_{t+1} = X_t + A\mathcal{I}_{t+1}, \tag{1.3}$$

where $\mathcal{I}_{t+1}$ is a $N \times K$ random matrix representing which content is relative to the broadcast of agent j at time $t+1$. The matrix $\mathcal{I}_{t+1}$ has entries

$$\mathcal{I}_{t+1}^{i,k} = \mathbb{I}_{\{I_{t+1}^i = k\}}.$$

Figure 1.1 illustrates an example of both the broadcasting and updating mechanisms.

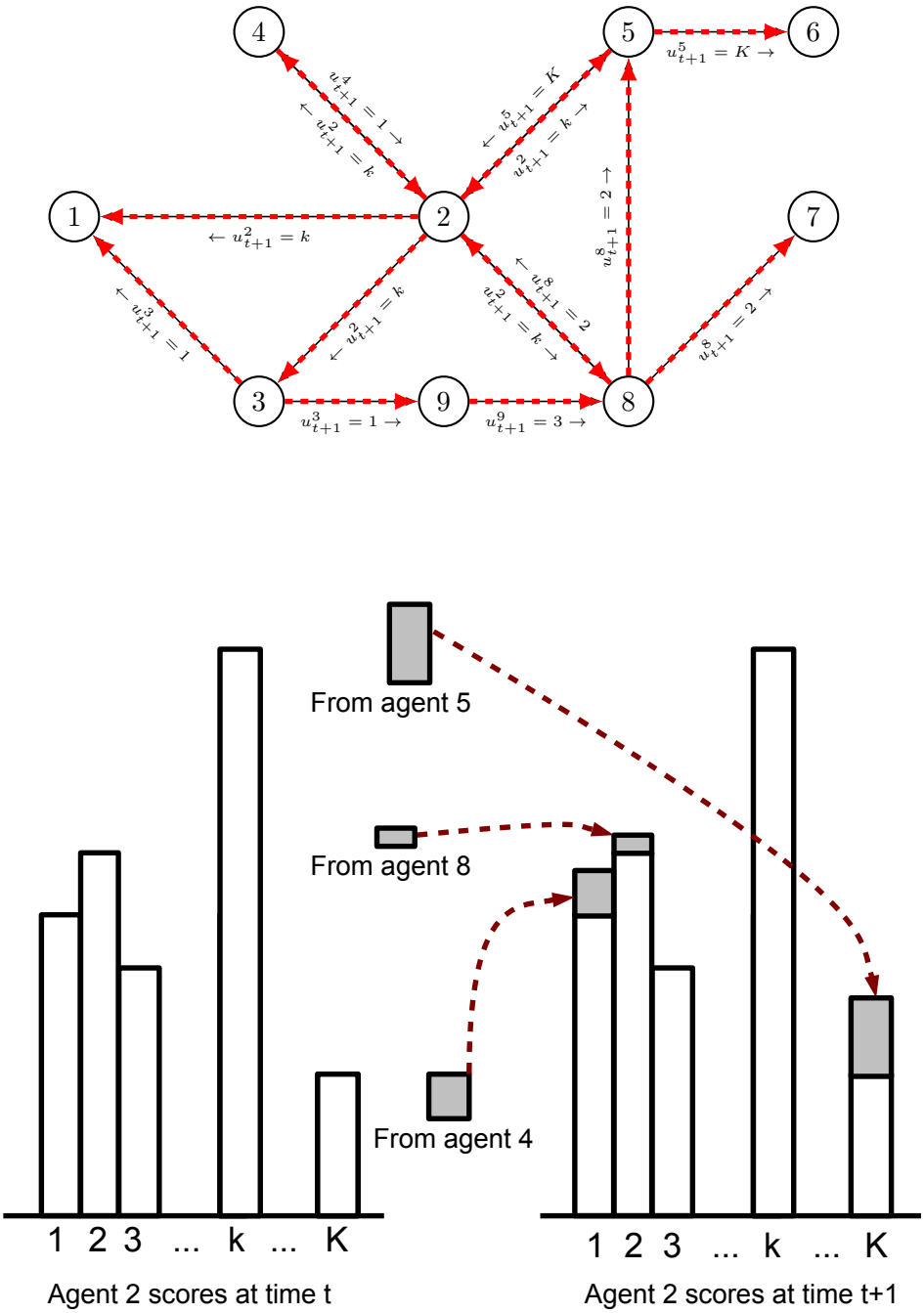

Figure 1.1: At $t+1$ each agent chooses a content from the distribution (1.1) and broadcasts it to her neighbors (broadcasting step), then agents update their scores with the received information, following Eqn. (1.2) (updating step).

1.2.3 Discussion of the model

As previously discussed, our opinion dynamics model has two building steps: a *broadcasting* step, in which agents select the contents relative to their broadcast and transmit the information

to neighbors, and an *updating* step, in which agents retrieve the transmitted information from their neighbors and update their scores according to the contents relative to each broadcast. These steps are, however, independent from each other: the broadcasting step takes into consideration only the agents' scores, by means of the random sampling procedure, whereas the updating step relies exclusively on agents' influences, which affect how strongly agents interpret the information received during the broadcast step.

The random sampling procedure is performed using the softmax function with parameter β [28]. The softmax parameter β impacts the choice of the content during the random sampling: a small value of β corresponds to a uniform choice over the K contents, and as β grows, the sampling becomes more biased towards contents with larger scores, i.e., agents start to broadcast more often contents they appreciate the most.

It is worth mentioning that qualitatively the scores X_t and the normalized scores P_t represent the same thing: agents appreciation of contents. The difference is that X_t represents absolute scores of agents while P_t represents relative scores of agents.

1.2.4 Assumptions and main result

1.2.4.1 Assumptions

In order to prove our main convergence result, we make the following assumption throughout the chapter:

Assumption 1. *(i)* $\min_i D_{ii} = \min_i \sum_j A_{ij} > 0$,

(ii) $\min_i \sum_k X_0^{i,k} > 0$.

Assumption D.1(i) implies that every agent is influenced, and thus agents scores are updated at each step of the opinion dynamics algorithm. The degree matrix D then satisfies $D_{ii} = \sum_j A_{ij} > 0$ for every $i \in V$, which allows the definition of the inverse matrix D^{-1}.

Assumption D.1(ii) implies that each agent has an initial score, which simply serves to avoid a different starting rule for the opinion dynamics algorithm. $X_0^{i,k}$ represents the initial opinion of agent i about content k; the bigger this opinion the harder is for agent i to change it during the opinion dynamics algorithm.

1.2.4.2 Main result

The main result of this chapter provides the almost sure convergence of agents normalized scores P_t subject to the opinion dynamics (1.2), under assumption D.1. The next section is dedicated to prove the following statement:

Theorem 1. *Let $P_t \in \Delta_K^N$ be the agents normalized scores defined as $P_t = D(X_t)^{-1}X_t$, where the agents scores X_t follow the updating Eqn. (1.2).*

Under assumption D.1, we have that $P_t \to \mathcal{F}_\beta^x$ almost surely when $t \to \infty$, where

$$\mathcal{F}_\beta^x = \{x \in \Delta_K^N \mid x = D^{-1}Af_\beta(x)\}.$$

1.3 Convergence analysis and proof of theorem 1

We provide in this section a rigorous proof of theorem 1, which dictates the convergence of the agents normalized scores P_t defined in section 1.2. The proof relies on the description of the evolution of the normalized scores P_t as a stochastic approximation algorithm [23, 185], which allows the use of the so-called ODE method (see for example [41]) to ensure the desired convergence result.

1.3.1 Tools necessary for convergence

In order to fully understand the concepts used during the proof of theorem 1, we provide a quick introduction on stochastic approximation algorithms, the ODE method and Lyapunov functions. The interested reader is directed to [23, 185] for detailed tutorials.

1.3.1.1 Stochastic approximation algorithms

We say that the sequence of random matrices[1] $(S_t)_{t\in\mathbb{N}} \in \mathcal{M}_{N\times K}(\mathbb{R})$ is a stochastic approximation algorithm if S_t satisfies the following recursive equation

$$S_{t+1} = S_t + \rho_{t+1}\Big(g(S_t) + M_{t+1} + r_{t+1}\Big), \tag{1.4}$$

where

- the step size $\rho_t \in \mathbb{R}_+$ satisfies $\rho_t \to 0$ and $\sum_t \rho_t = \infty$,
- $g : \mathcal{M}_{N\times K}(\mathbb{R}) \to \mathcal{M}_{N\times K}(\mathbb{R})$ is a continuous function,
- $M_{t+1} \in \mathcal{M}_{N\times K}(\mathbb{R})$ is a martingale difference, i.e., $\mathbb{E}[M_{t+1}|\sigma(S_s, M_s, r_s,\ s \le t)] = 0$, and
- the remainder term $r_{t+1} \in \mathcal{M}_{N\times K}(\mathbb{R})$ satisfies $r_{t+1} \to 0$ almost surely.

Stochastic approximation algorithms can be seen as the random counterpart of the Euler discretisation of the ordinary differential equation (ODE)

$$\dot{s} = g(s), \quad s_0 \in \mathcal{M}_{N\times K}(\mathbb{R}), \tag{1.5}$$

and under mild assumptions possess the same asymptotic behavior as the semiflow induced by it (see [23, 185]), as discussed next.

Let us define by $\overline{S} : \mathbb{R}_+ \to \mathcal{M}_{N\times K}(\mathbb{R})$ the continuous time affine interpolation of S_t, such that $\tau_0 = 0$, $\tau_t = \sum_{i=1}^{t} \rho_i$ and

$$\overline{S}(\tau_t + s) = S_t + s\frac{S_{t+1} - S_t}{\tau_{t+1} - \tau_t} \tag{1.6}$$

for all $t \in \mathbb{N}$ and $0 \le s < \rho_{t+1}$, and let $\Phi^g : \mathbb{R}_+ \times \mathcal{M}_{N\times K}(\mathbb{R}) \to \mathcal{M}_{N\times K}(\mathbb{R})$ be the semiflow induced by ODE (1.5).

When the function g is bounded and Lipschitz continuous (or the stochastic approximation algorithm is bounded almost surely), the noise in the stochastic approximation algorithm has bounded variance, and both the step size and remainder term decrease sufficiently fast, we have that the process $\overline{S}$ shadows in every interval $[t, T + t]$ the semiflow Φ^g originated in $\overline{S}(t)$, when t is large enough. This is due to the next lemma, consequence of propositions 4.1 and 4.2 of [23]:

1. We use for simplicity random matrices, but the reader may see [23] for a more general definition using metric spaces.

Lemma 1. *Let S_t be the stochastic approximation algorithm defined by Eqn. (1.4) and $\overline{S}$ its continuous time affine interpolation.*

If g is bounded and Lipschitz continuous (or if $\sup_t ||S_t|| < \infty$ almost surely), $\sup_t \mathbb{E}[||M_{t+1}||^2] < \infty$, $\sum_t \gamma_t^2 < \infty$ and $\sum_t \rho_t ||r_t|| < \infty$ almost surely, we have that

$$\lim_{t\to\infty} \sup_{0\leq h\leq T} ||\overline{S}(t+h) - \Phi^g(h, \overline{S}(t))|| = 0 \tag{1.7}$$

for any $T > 0$, where Φ^g is the semiflow induced by the limit ODE (1.5).

Definition 1. *We say that the continuous time affine interpolation $\overline{S}$ is an Asymptotic Pseudotrajectory (APT) of the semiflow Φ^g induced by ODE (1.5) (see [23]) if it satisfies Eqn. (1.7) for the semiflow Φ^g.*

1.3.1.2 The ODE method

After introducing the concepts of stochastic approximation algorithms and asymptotic pseudotrajectories of semiflows, we are ready to discuss the ODE method, which is one of the basic tools used to proving convergence of stochastic approximation algorithms.

The ODE method works as follows:

1. First, one retrieves the limit ODE (1.5) from the stochastic approximation algorithm S_t given by Eqn. (1.4).

2. Second, one proves that $\overline{S}$, the continuous time affine interpolation (1.6), is an asymptotic pseudotrajectory of Φ^g, the semiflow associated with ODE (1.5), i.e., $\overline{S}$ satisfies Eqn. (1.7).

3. Finally, one proves that Φ^g converges towards some limit set, which in turn implies the convergence of the stochastic approximation algorithm S_t towards the same limit set under mild assumptions.

1.3.1.3 Lyapunov functions

A crucial step of the ODE method consists in proving that the semiflow Φ^g induced by the limit ODE (1.5) converges to a limit set, which can be achieved for example through the construction of a Lyapunov function (see [23] for instance).

Definition 2. *Let $\Lambda \subset \mathcal{M}_{N\times K}(\mathbb{R})$ be a compact invariant set of the semiflow Φ^g induced by the limit ODE (1.5). We say that a continuous function $V : \mathcal{M}_{N\times K}(\mathbb{R}) \to \mathbb{R}$ is a Lyapunov function for Λ if*

- *The function $t \to V(\Phi^g(t, x))$ is strictly decreasing if $x \in \mathcal{M}_{N\times K}(\mathbb{R}) \setminus \Lambda$.*
- *The function $t \to V(\Phi^g(t, x))$ is constant if $x \in \Lambda$.*

1.3.2 Sketch of proof

We start the proof of theorem 1 by providing some insights, all arguments are made rigorous in the remainder of the section.

The proof is performed in several steps, following the ODE method of subsubsection 1.3.1.2:

(i) We study the evolution of normalized scores $P_t = D(X_t)^{-1}X_t$ under the updates (1.2). We show that the normalized scores P_t can be represented as a stochastic approximation algorithm [23, 185] and that they admit a stochastic approximation algorithm decomposition (Y_t, W_t) that satisfies $P_t = D^{-1}AY_t + W_t$ and where its continuous time affine interpolation $(\overline{Y}, \overline{W})$ is an Asymptotic PseudoTrajectory of the semiflow Φ associated with the limit ODE

$$\begin{cases} \dot{y} &= f_\beta(D^{-1}Ay + w) - y \\ \dot{w} &= -w. \end{cases} \tag{1.8}$$

(ii) We study the limit behavior of the semiflow induced by ODE (1.8), from which the proof of theorem 1 is a direct consequence of proposition 6.4 in [23]. To do so, we proceed as follows:

(ii-a) We derive a Lyapunov function V_β for the semiflow Φ, following the ideas in [150], which in turn implies the convergence of Φ to the set $(\mathcal{F}_\beta^y, 0)$, where $\mathcal{F}_\beta^y = \{y \in \Delta_K^N \mid y = f_\beta(D^{-1}Ay)\}$, and the almost sure convergence of (Y_t, W_t) to the set $(\mathcal{F}_\beta^y, 0)$.

(ii-b) Due to the almost sure convergence of (Y_t, W_t) to $(\mathcal{F}_\beta^y, 0)$, we retrieve the almost sure convergence of $P_t = D^{-1}AY_t + W_t$ to the set $D^{-1}A\mathcal{F}_\beta^y$, which we show is in fact equal to the set $\mathcal{F}_\beta^x = \{x \in \Delta_K^N \mid x = D^{-1}Af_\beta(x)\}$.

1.3.3 The opinion dynamics algorithm as a stochastic approximation algorithm

As previously mentioned, theorem 1 provides the limit behavior of normalized scores P_t under the updates given by Eqn. (1.2), which can be written in matrix form as Eqn. (1.3).

We begin the proof of theorem 1, following the steps of the ODE method detailed in subsubsection 1.3.1.2, with a lemma describing P_t as a stochastic approximation algorithm.

Lemma 2. *We have that the normalized scores P_t satisfy the following stochastic approximation algorithm for $t \in \mathbb{N}$*

$$P_{t+1} = P_t + \frac{1}{t+1}\Big(D^{-1}Af_\beta(P_t) - P_t + \zeta_{t+1} + \eta_{t+1}\Big), \tag{1.9}$$

where ζ_{t+1} is a bounded martingale difference, i.e., $\mathbb{E}[\zeta_{t+1}|\mathcal{G}_t] = 0$, and η_{t+1} is a bounded random matrix satisfying $\sum_t \frac{1}{t+1}||\eta_{t+1}|| < \infty$.

Proof. By the definition of the normalized score matrix P_t, one has that Eqn. (1.3) can be written as

$$D(X_{t+1})P_{t+1} = D(X_t)P_t + A\mathcal{I}_{t+1},$$

where P_t is the $N \times K$ matrix with agents normalized scores $P_t^{i,k}$ at time t and $\mathcal{I}_{t+1}$ is a random matrix accounting for the updating of the algorithm.

The matrix $\mathcal{I}_{t+1}$ has entries $\mathcal{I}_{t+1}^{i,k} = \mathbb{I}_{\{I_{t+1}^i = k\}}$ and we clearly have by Eqn. (1.1) that $\mathbb{E}[\mathcal{I}_{t+1}^{i,k}|\mathcal{G}_t] = \mathbb{P}(I_{t+1}^i = k|\mathcal{G}_t) = f_\beta^{i,k}(P_t)$, i.e.,

$$\mathbb{E}[\mathcal{I}_{t+1}|\mathcal{G}_t] = f_\beta(P_t),$$

hence $\overline{\zeta}_{t+1} = A\mathcal{I}_{t+1} - Af_\beta(P_t)$ satisfies $\mathbb{E}[\overline{\zeta}_{t+1}|\mathcal{G}_t] = 0$, i.e., $\overline{\zeta}_{t+1}$ is a martingale difference.

Eqn. (1.3) resolves to

$$D(X_{t+1})P_{t+1} = D(X_t)P_t + A\mathcal{I}_{t+1} = D(X_t)P_t + \overline{\zeta}_{t+1} + Af_\beta(P_t),$$

which implies by subtracting $D(X_{t+1})P_t$ from both sides

$$D(X_{t+1})(P_{t+1} - P_t) = (D(X_t) - D(X_{t+1}))P_t + Af_\beta(P_t) + \overline{\zeta}_{t+1}.$$

We have for every $i \in V$

$$D(X_{t+1})_{ii} = (X_{t+1}1)_i = (X_t 1 + A\mathcal{I}_{t+1}1)_i = (X_t 1)_i + (A1)_i = D(X_t)_{ii} + D_{ii},$$

which implies that

$$\begin{aligned} P_{t+1} &= P_t + D(X_{t+1})^{-1}\Big(- DP_t + Af_\beta(P_t) + \overline{\zeta}_{t+1}\Big) \\ &= P_t + \frac{1}{t+1}\Big(D^{-1}Af_\beta(P_t) - P_t + \zeta_{t+1} + \eta_{t+1}\Big), \end{aligned}$$

where $\zeta_{t+1} = D^{-1}\overline{\zeta}_{t+1}$ is a bounded martingale difference and the remainder term

$$\eta_{t+1} = \Big((t+1)D(X_{t+1})^{-1} - D^{-1}\Big)(Af_\beta(P_t) - DP_t + \overline{\zeta}_{t+1})$$

satisfies, for constants $L, L' > 0$,

$$\sum_t \frac{1}{t+1}||\eta_{t+1}|| \leq L\sum_t \frac{1}{t+1}||(D + \frac{D(X_0)}{t+1})^{-1} - D^{-1}|| \leq L'\sum_t \frac{1}{(t+1)^2} < \infty$$

because $D(X_t) = tD + D(X_0)$, and f_β and $\overline{\zeta}_{t+1}$ are bounded. □

1.3.4 Decomposition of preferences

After deriving a stochastic approximation algorithm for the evolution of preferences P_t, we continue the ODE method by proving that the semiflow induced by its limit ODE converges, however the limit ODE of the stochastic approximation algorithm satisfied by the agents preferences P_t does not have any special structure that allows us to prove its convergence. Nevertheless, by decomposing P_t into two parts Y_t and W_t, we are able to derive a new stochastic approximation algorithm possessing a structure allowing us to prove its convergence.

We start thus the convergence analysis of the preferences P_t by decomposing Eqn. (1.9) into a new stochastic approximation algorithm (Y_t, W_t) composed of two parts Y_t and W_t; we are able to prove that the part W_t converges to zero almost surely, whereas one has that the part Y_t bares resemblance to the stochastic fictitious play studied in [150]. This step is extremely important since it allows the eventual derivation of a Lyapunov function V_β for the semiflow induced by the limit ODE associated with the couple (Y_t, W_t).

We begin the decomposition with two auxiliary lemmas:

Lemma 3. *We have that*

$$\sup_{x \in \mathcal{M}_{N\times K}(\mathbb{R})} \max_{(i,k),(j,c)} |\partial_{(j,c)} f_\beta^{i,k}(x)| \leq \beta,$$

which implies that $f_\beta : \mathcal{M}_{N\times K}(\mathbb{R}) \to \Delta_K^N$ has bounded derivative and is Lipschitz continuous.

Proof. The conclusion easily follows from the calculation of the derivatives of f_β

$$\partial_{(j,c)} f_\beta^{i,k}(x) = \begin{cases} 0 \text{ if } j \neq i \\ \beta\Big(f_\beta^{i,k}(x) - (f_\beta^{i,k}(x))^2\Big) \text{ if } j = i,\ c = k \\ -\beta f_\beta^{i,k}(x) f_\beta^{i,c}(x) \text{ if } j = i,\ c \neq k, \end{cases}$$

and the fact that $0 \leq f_\beta^{i,k}(x) \leq 1$ for all (i,k). □

Lemma 4. *Define* $B_{2\sqrt{N}} = \{x \in \mathcal{M}_{N\times K}(\mathbb{R}) \mid ||x|| \leq 2\sqrt{N}\}$. *There exists a* $0 < \delta \leq \frac{1}{K}$ *such that*

$$\inf_{(z,w)\in\Delta_K^N \times B_{2\sqrt{N}}} \min_{(i,k)} f_\beta^{i,k}(D^{-1}Az + w) \geq \delta.$$

Proof. The result easily stems from the fact that $f_\beta^{i,k}(11^T/K) = \frac{1}{K}$, the compactness of the set $\{D^{-1}Az+w \mid (z,w) \in \Delta_K^N \times B_{2\sqrt{N}}\} \subset \mathcal{M}_{N\times K}(\mathbb{R})$ and the continuity of f_β, since $f_\beta^{i,k}(D^{-1}Az+w) > 0$ for all $(z,w) \in \Delta_K^N \times B_{2\sqrt{N}}$. □

Now, we state the main decomposition lemma:

Lemma 5. *Let* P_t *be the stochastic approximation algorithm* (1.9). *Define by* (Y_t, W_t) *the following stochastic approximation algorithm*

$$\begin{cases} Y_{t+1} = Y_t + \frac{1}{t+1}\Big(f_\beta(D^{-1}AY_t + W_t) - Y_t\Big) \\ W_{t+1} = W_t + \frac{1}{t+1}\Big(- W_t + \zeta_{t+1} + \eta_{t+1}\Big), \end{cases} \tag{1.10}$$

with $Y_0 = \frac{11^T}{K}$, $W_0 = P_0 - \frac{11^T}{K}$, *and where* ζ_{t+1} *and* η_{t+1} *are defined in lemma 2.*

We have that

(i) $P_t = D^{-1}AY_t + W_t$ *for all* $t \in \mathbb{N}$.

(ii) Let $\delta > 0$ *be the constant defined in lemma 4. Then* $Y_t \in \{y \in \Delta_K^N \mid y^{i,k} \geq \delta\},\ \forall (i,k)\}$ *for all* $t \in \mathbb{N}$ *and* $\sup_t ||W_t|| \leq 2\sqrt{N}$.

(iii) The continuous time affine interpolation $(\overline{Y}, \overline{W}) : \mathbb{R}_+ \to \mathcal{M}_{N\times K}(\mathbb{R}) \times \mathcal{M}_{N\times K}(\mathbb{R})$ *of* (Y_t, W_t) *is an Asymptotic Pseudotrajectory of* $\Phi : \mathbb{R} \times \Big(\mathcal{M}_{N\times K}(\mathbb{R}) \times \mathcal{M}_{N\times K}(\mathbb{R})\Big) \to \mathcal{M}_{N\times K}(\mathbb{R}) \times \mathcal{M}_{N\times K}(\mathbb{R})$, *the semiflow induced by the following ODE*

$$\begin{cases} \dot{y} &= f_\beta(D^{-1}Ay + w) - y \\ \dot{w} &= -w. \end{cases} \tag{1.11}$$

Proof. (i) Let us define $Q_t = D^{-1}AY_t + W_t$. We must prove that $Q_t = P_t$, which we do by induction in t. The result is clearly true for $t = 0$ since $D^{-1}A\frac{11^T}{K} = \frac{11^T}{K}$ implies

$$P_0 = \frac{11^T}{K} + (P_0 - \frac{11^T}{K}) = D^{-1}AY_0 + W_0 = Q_0.$$

Let us assume the result is true for t, i.e., $P_t = Q_t$. Thus

$$\begin{aligned} Q_{t+1} &= D^{-1}AY_{t+1} + W_{t+1} \\ &= D^{-1}AY_t + W_t + \frac{1}{t+1}\Big(D^{-1}Af_\beta(D^{-1}AY_t + W_t) - D^{-1}AY_t - W_t + \zeta_{t+1} + \eta_{t+1}\Big) \\ &= P_t + \frac{1}{t+1}\Big(D^{-1}Af_\beta(P_t) - P_t + \zeta_{t+1} + \eta_{t+1}\Big) \\ &= P_{t+1} \end{aligned}$$

since P_t satisfies Eqn. (1.9), which proves item (i).

(ii) First of all, we prove by induction in t that $Y_t^{i,k} \geq \delta$ for all $t \in \mathbb{N}$: the base case $Y_0^{i,k} = \frac{1}{K} \geq \delta$ stems from the definition of δ in lemma 4. Suppose that $Y_t^{i,k} \geq \delta$, then, since $P_t \in \Delta_K^N$ by construction and $f_\beta^{i,k}(P_t) \geq \delta$ for all P_t by lemma 4, we have that

$$\begin{aligned} Y_{t+1}^{i,k} &= Y_t^{i,k} + \frac{1}{t+1}\Big(f_\beta^{i,k}(D^{-1}AY_t + W_t) - Y_t^{i,k}\Big) = Y_t^{i,k}(1 - \frac{1}{t+1}) + \frac{1}{t+1}f_\beta^{i,k}(P_t) \\ &\geq Y_t^{i,k}(1 - \frac{1}{t+1}) + \frac{\delta}{(t+1)} \geq \delta, \end{aligned}$$

which completes the induction.

Let us now prove by induction in t that $Y_t 1 = 1$ for all $t \geq 0$: we have that $Y_0 1 = 1$, which is the base case. Let us assume now that $Y_t 1 = 1$, then

$$Y_{t+1}1 = Y_t 1 + \frac{1}{t+1}\Big(f_\beta(D^{-1}AY_t + W_t) - Y_t\Big)1 = Y_t 1 = 1,$$

since $f_\beta(z) \in \Delta_K^N$ for all $z \in \mathcal{M}_{N\times K}(\mathbb{R})$, which completes the induction.

Since $W_t = P_t - D^{-1}AY_t$ by item (i), $P_t \in \Delta_K^N$ for all $t \in \mathbb{N}$ by construction, we just proved that $Y_t \in \Delta_K^N$ for all $t \in \mathbb{N}$ and Δ_K^N is invariant by $D^{-1}A$, the fact that $\sup_{z\in\Delta_K^N} ||z|| \leq \sqrt{N}$ concludes item (ii).

(iii) First of all, we have that the semiflow $\Phi : \mathbb{R} \times \Big(\mathcal{M}_{N\times K}(\mathbb{R}) \times \mathcal{M}_{N\times K}(\mathbb{R})\Big) \to \mathcal{M}_{N\times K}(\mathbb{R}) \times \mathcal{M}_{N\times K}(\mathbb{R})$ is globally defined and possesses unique trajectories since f_β is smooth and Lipschitz continuous by lemma 3.

In view of Eqn. (1.10), since $\sup_{t\in\mathbb{N}}(||Y_t|| + ||W_t||) < \infty$ by item (ii), f_β is bounded and Lipschitz continuous, $\sum_t \frac{1}{(t+1)^2} < \infty$, ζ_{t+1} is a bounded martingale and $\sum_t \frac{1}{t+1}||\eta_{t+1}|| < \infty$, the proof follows from lemma 1.

□

1.3.5 Lyapunov function for the limit ODE (1.11)

We now proceed to proving that the semiflow Φ induced by the limit ODE (1.11) converges. This is achieved by the construction of a Lyapunov function V_β for the invariant set $(\mathcal{F}_\beta^y, 0) \in \Delta_K^N \times \mathcal{M}_{N\times K}(\mathbb{R})$, where $\mathcal{F}_\beta^y = \{y \in \Delta_K^N \mid y = f_\beta(D^{-1}Ay)\}$.

The Lyapunov function is composed of two parts, each one related to a term of the decomposition (Y_t, W_t) defined in lemma 5. The part related to Y_t is exactly the Lyapunov function constructed

in [150] for potential games in stochastic fictitious play, and the part related to W_t is proportional to the norm of the autonomous part of ODE (1.11), with the proportionality constant taking into consideration the derivative of the entropy function $\mathcal{H}_\beta$.

However, since the norm of the derivative of $\mathcal{H}_\beta$ goes to ∞ at the boundary $\partial\Delta_K^N = \{y \in \Delta_K^N \mid \exists(i,k) \text{ such that } y^{i,k} = 0\}$, we must restrict the semiflow Φ induced by the limit ODE (1.11) to an invariant compact subset $\mathcal{K}$ of $\mathring{\Delta}_K^N$ such that $Y_t \in \mathcal{K}$ for all $t \in \mathbb{N}$.

We begin the construction of the Lyapunov function with some auxiliary lemmas:

Lemma 6. *We have that $\mathring{\Delta}_K^N = \{x \in \mathcal{M}_{N\times K}(\mathbb{R}) \mid x1 = 1 \text{ and } x_{i,k} > 0,\ \forall(i,k)\}$ is a smooth manifold of dimension $N \times (K-1)$ without boundary.*

Moreover, the tangent space at $x \in \mathring{\Delta}_K^N$ is given by

$$T_x\mathring{\Delta}_K^N = \{\lambda \in \mathcal{M}_{N\times K}(\mathbb{R}) \mid \lambda 1 = 0\}.$$

Proof. The result stems from the fact that $\mathring{\Delta}_K^N$ is an open subset, for the induced topology, of the affine subspace $\mathcal{S} = \{m \in \mathcal{M}_{N\times K}(\mathbb{R}) \mid m1 = 1\}$, with associated vector space $\mathcal{V} = \{\lambda \in \mathcal{M}_{N\times K}(\mathbb{R}) \mid \lambda 1 = 0\}$. □

Lemma 7. *Let (y_t, w_t) be the solution of ODE (1.11) with $y_0 \in \mathring{\Delta}_K^N$, given by lemma 5. There exists a nonempty maximal interval $J = [0, t^*)$ such that $y_t \in \mathring{\Delta}_K^N = \{x \in \mathcal{M}_{N\times K}(\mathbb{R}) \mid x1 = 1 \text{ and } x_{i,k} > 0,\ \forall(i,k)\}$ for all $t \in J$.*

Proof. Define $t^* = \sup\{s \geq 0 \mid y_s \in \mathring{\Delta}_K^N \text{ and } \nexists u \in [0,s] \text{ such that } y_u \notin \mathring{\Delta}_K^N\}$ as the supremum of the times for which y_t remains in $\mathring{\Delta}_K^N$ before it exits for the first time.

We have for $(y, w) \in \mathring{\Delta}_K^N \times \mathcal{M}_{N\times K}(\mathbb{R})$ that

$$\Big(f_\beta(D^{-1}Ay + w) - y\Big)1 = f_\beta(D^{-1}Ay + w)1 - y1 = 1 - 1 = 0$$

since $f : \Delta_K^N \to \Delta_K^N$, which implies that for $y_t \in \mathring{\Delta}_K^N$ we have $\dot{y}_t = f_\beta(D^{-1}Ay_t + w_t) - y_t \in T_{y_t}\mathring{\Delta}_K^N$ by lemma 6.

Since $y_0 \in \mathring{\Delta}_K^N$, $\mathring{\Delta}_K^N$ is a smooth manifold by lemma 6, f_β is smooth and $\dot{y}_t \in T_{y_t}\mathring{\Delta}_K^N$ for $y_t \in \mathring{\Delta}_K^N$, we have by standard theory of ODEs in manifolds that $y_t \in \mathring{\Delta}_K^N$ for all $0 \leq t < t^*$, and that $t^* > 0$. □

Lemma 8. *We have that for every $y \in \mathcal{F}_\beta^y$ there exists a vector $c(y) \in \mathbb{R}^N$ such that*

$$\nabla_y V_\beta(y, w) = c(y)1^T, \tag{1.12}$$

where the Lyapunov function V_β is defined in lemma 10.

Proof. This lemma is simply a particular case of lemma A.1 in [150]. However we provide a proof for the sake of completeness. We easily have that

$$\begin{aligned}\partial_{y^{i,k}} V_\beta(y, w) &= -\frac{1}{2}\Big(\sum_j A_{ij}y^{j,k} + \sum_j A_{ji}y^{j,k}\Big) + \frac{D_{ii}}{\beta}\log(y^{i,k}) + \frac{D_{ii}}{\beta} \\ &= \frac{D_{ii}}{\beta}\Big(\log(y^{i,k}) - \beta(D^{-1}Ay)_{i,k} + 1\Big)\end{aligned}$$

by the symmetry of the adjacency matrix A.

Since $y \in \mathcal{F}_\beta^y$, we have that

$$\log(y^{i,k}) = \beta(D^{-1}Ay)_{i,k} - \log(\sum_{k'} e^{\beta(D^{-1}Ay)_{i,k'}}),$$

which implies that

$$\partial_{y^{i,k}} V_\beta(y,w) = \frac{D_{ii}}{\beta}\left(1 - \log(\sum_{k'} e^{\beta(D^{-1}Ay)_{i,k'}})\right) = c^i(y)$$

and concludes the proof. □

Lemma 9. *Let $(\hat{F}^i)_{i\in V}$ be the auxiliary functions defined in lemma 10. We have for every $y \in \Delta_K^N$ that for all $i \in V$*

$$\langle f^i(D^{-1}Ay) - y^i, \hat{F}^i(y)\rangle \geq 0.$$

Proof. This lemma is simply a particular case of lemma A.2 in [150]. However, we provide a proof for the sake of completeness.

First of all, we have that $\nabla_k \mathcal{H}_\beta(x) = \frac{1}{\beta}(\log x^k + 1)$, which implies that

$$\begin{aligned}\nabla_k \mathcal{H}_\beta(x)_{|x=f_\beta^i(D^{-1}Ay)} &= (D^{-1}Ay)_{i,k} + \frac{1}{\beta}(1 - \log(\sum_{k'} e^{\beta(D^{-1}Ay)_{i,k'}})) \\ &= (D^{-1}Ay)_{i,k} + a^i(y),\end{aligned}$$

with $a^i(y) \in \mathbb{R}$, which can be written in a more concise form

$$\nabla \mathcal{H}_\beta(x)_{|x=f_\beta^i(D^{-1}Ay)} = (D^{-1}Ay)_i + a^i(y)1.$$

Moreover, since $y \in \Delta_K^N$ and $f_\beta(D^{-1}Ay) \in \Delta_K^N$, we have that

$$\langle a^i(y)1, f_\beta^i(D^{-1}Ay) - y^i\rangle = a^i(y)\big(1-1\big) = 0.$$

Thus,

$$\begin{aligned}&\langle \hat{F}^i(y), f^i(D^{-1}Ay) - y^i\rangle = \langle (D^{-1}Ay)_i - \nabla\mathcal{H}_\beta(y^i), f^i(D^{-1}Ay) - y^i\rangle \\ &= \langle \nabla\mathcal{H}_\beta(x)_{|x=f_\beta^i(D^{-1}Ay)} - \nabla\mathcal{H}_\beta(y^i), f^i(D^{-1}Ay) - y^i\rangle - \langle a^i(y)1, f^i(D^{-1}Ay) - y^i\rangle \\ &= \langle \nabla\mathcal{H}_\beta(x)_{|x=f_\beta^i(D^{-1}Ay)} - \nabla\mathcal{H}_\beta(y^i), f^i(D^{-1}Ay) - y^i\rangle \geq 0\end{aligned}$$

since the entropy function $\mathcal{H}_\beta$ is convex. □

We now state the lemma constructing the Lyapunov function:

Lemma 10. *Let $\delta > 0$ be the constant defined in lemma 4, $\mathcal{K} = \{y \in \Delta_K^N \mid y^{i,k} \geq \frac{\delta}{4},\ \forall (i,k)\} \subset \mathring{\Delta}_K^N$, $B_{2\sqrt{N}} = \{x \in \mathcal{M}_{N\times K}(\mathbb{R}) \mid ||x|| \leq 2\sqrt{N}\}$, and define the functions $\hat{F} : \mathcal{K} \to \mathcal{M}_{N\times K}(\mathbb{R})$ and $V : \mathcal{K} \times B_{2\sqrt{N}} \to \mathbb{R}$ by*

$$\hat{F}^i(y) = \sum_j D_{ii}^{-1} A_{ij} y^j - \nabla\mathcal{H}_\beta(y^i), \quad \forall i \in V,$$

and

$$V_\beta(y,w) = L||w|| - \sum_i D_{ii}\left(\frac{1}{2}\sum_j D_{ii}^{-1}A_{ij}\langle y^i, y^j\rangle - \mathcal{H}_\beta(y^i)\right),$$

where

$$L = 2\big(\sup_{z\in\mathcal{M}_{N\times K}(\mathbb{R})} ||\nabla f_\beta(z)||\big).\big(\sup_{x\in\mathcal{K}}\sum_i D_{ii}||\hat{F}^i(x)||\big).$$

We have that

(i) the compact set $\mathcal{K}\times B_{2\sqrt{N}}$ is an invariant set of the semiflow Φ induced by the limit ODE (1.11).

(ii) $(\mathcal{F}^y_\beta, 0)\subset \mathcal{K}\times B_{2\sqrt{N}}$ and $(\overline{Y}(t),\overline{W}(t))\in\mathcal{K}\times B_{2\sqrt{N}}$ for all $t\in\mathbb{R}_+$.

(iii) V_β is a Lyapunov function for the set $(\mathcal{F}^y_\beta,0)$ with respect to $\Phi_{|\mathcal{K}\times B_{2\sqrt{N}}}$, the semiflow induced by the limit ODE (1.11) and restricted to $\mathcal{K}\times B_{2\sqrt{N}}$.

(iv) $V_\beta(\mathcal{F}^y_\beta,0)=\bigcup_{z\in\mathcal{F}^y_\beta} V_\beta(z,0)$ has empty interior.

Proof. (i) Let (y_t,w_t) be the unique solution of ODE (1.11) such that $(y_0,w_0)\in\mathcal{K}\times B_{2\sqrt{N}}$, i.e., $(y_t,w_t)=\Phi(t,(y_0,w_0))$.

First of all, since $\dot{w}=-w$, we trivially have that the function $||w_t||^2$ satisfies $\frac{d}{dt}||w_t||^2 = -2||w_t||^2\leq 0$, which implies that $w_t\in B_{2\sqrt{N}}$ for all $t\geq 0$.

Take now $w_0\in B_{2\sqrt{N}}$ and $y_0\in\partial\mathcal{K}=\{z\in\Delta_K^N \mid \exists(i,k) \text{ such that } z^{i,k}=\frac{\delta}{4}\}$ and let (i_k,k_n) be indices such that $y_0^{i_n,k_n}=\frac{\delta}{4}$ and that $y_0^{i,k}>\frac{\delta}{4}$ for all $(i,k)\notin\bigcup_n(i_n,k_n)$.

Define now, for $\epsilon>0$, the set $B_\epsilon(y_0,w_0)=\{(y,w)\in\Delta_K^N\times B_{2\sqrt{N}} \mid ||y-y_0||+||w-w_0||\leq\epsilon\}$. We clearly have that there exists a $\epsilon>0$ such that $\max_n y^{i_n,k_n}\leq\frac{\delta}{2}$ for all $(y,w)\in B_\epsilon(y_0,w_0)$. Furthermore, we also have by lemma 4 that $\min_{(i,k)} f_\beta^{i,k}(D^{-1}Ay+w)\geq\delta$.

Moreover, by the continuity of y_t and by lemma 7, there exists a nonempty interval $[0,t^*)$ such that $(y_t,w_t)\in B_\epsilon(y_0,w_0)$ and $\min_{(i,k)\notin\bigcup_n(i_n,k_n)} y_t^{i,k}>\frac{\delta}{4}$ for all $t\in[0,t^*)$.

In addition, for every $t\in(0,t^*)$, the mean value theorem gives us

$$y_t = y_0+\int_0^t \dot{y}_u du = y_0+\int_0^t (f_\beta(D^{-1}Ay_u+w_u)-y_u)du.$$

Now, since $y_u^{i_n,k_n}\leq\frac{\delta}{2}$ and $f_\beta^{i_n,k_n}(D^{-1}Ay_u+w_u)\geq\delta$ for all $u\in[0,t]$, we have that

$$y_t^{i_n,k_n} = y_0^{i_n,k_n}+\int_0^t\left(f_\beta^{i_n,k_n}(D^{-1}Ay_u+w_u)-y_u^{i_n,k_n}\right)du \geq y_0^{i_n,k_n}+t\frac{\delta}{2} > y_0^{i_n,k_n}=\frac{\delta}{4}$$

and by consequence $y_t\in\mathcal{K}\setminus\partial\mathcal{K}$ for all $t\in(0,t^*)$. Since y_t can only exit the set $\mathcal{K}$ through an element of $\partial\mathcal{K}=\{y\in\Delta_K^N \mid \exists(i,k) \text{ such that } y^{i,k}=\frac{\delta}{4},\ \forall(i,k)\}$, we have the result for $y_0\in\partial\mathcal{K}$, which also implies the result for $y_0\in\mathcal{K}$ and concludes the proof of item (i).

(ii) Take $y\in\mathcal{F}^y_\beta$. Then by lemma 4 we have that $y^{i,k}=f_\beta^{i,k}(D^{-1}Ay)\geq\delta>\frac{\delta}{4}$, which proves that $\mathcal{F}^y_\beta\subset\mathcal{K}$. Hence item (ii) of lemma 5 and the convexity of $\mathcal{K}\times B_{2\sqrt{N}}$ conclude the proof of item (ii).

(iii) One simply must prove that the function $t\to V_\beta(y_t,w_t)$ is strictly decreasing for $(y_0,w_0)\notin(\mathcal{F}^y_\beta,0)$, since $(\mathcal{F}^y_\beta,0)$ is the equilibrium set of ODE (1.11), which clearly implies that $V_\beta(y_t,w_t)=V_\beta(y_0,w_0)$ if $(y_0,w_0)\in(\mathcal{F}^y_\beta,0)$.

First of all, since $\mathcal{K}\times B_{2\sqrt{N}}$ is invariant for the semiflow Φ by item (i) and $\sup_{x\in\mathcal{K}}\max_i||\nabla\mathcal{H}_\beta(x^i)|| < \infty$, the function $t\mapsto V_\beta(y_t,w_t)$ is well defined for all $t\geq 0$ when $(y_0,w_0)\in\mathcal{K}\times B_{2\sqrt{N}}$. Moreover, we have that $w_0\neq 0$ implies $w_t = w_0e^{-t}\neq 0$ for all $t\geq 0$, thus $t\mapsto V_\beta(y_t,w_t)$ is smooth in t when $w_0\neq 0$.

Using the symmetry of the adjacency matrix A we have that $\partial_{y^i}V_\beta(y,w) = -D_{ii}\hat{F}^i(y)$, thus following the proof of theorem 3.2 in [150], we have

$$\begin{aligned}\frac{d}{dt}V_\beta(y_t,w_t) &= L\langle\frac{w_t}{||w_t||},\dot{w}_t\rangle + \sum_i\langle\partial_{y^i}V_\beta(y_t,w_t),\dot{y}^i\rangle = -L||w_t|| - \sum_i D_{ii}\langle\hat{F}^i(y_t),\dot{y}^i\rangle\\ &= -L||w_t|| - \sum_i D_{ii}\langle\hat{F}^i(y_t), f^i_\beta(D^{-1}Ay_t) - y^i_t + \gamma^i_t\rangle,\end{aligned}$$

where $\gamma_t = f_\beta(D^{-1}Ay_t + w_t) - f_\beta(D^{-1}Ay_t)$.

Since $||\gamma_t||\leq\sup_{z\in\mathcal{M}_{N\times K}(\mathbb{R})}||\nabla f_\beta(z)||.||w_t||$ and $y_t\in\mathcal{K}$ by item (i), we have by the Cauchy-Schwarz inequality that

$$\sum_i D_{ii}|\langle\hat{F}^i(y_t),\gamma^i_t\rangle| \leq \sum_i D_{ii}||\hat{F}^i(y_t)||.||\gamma^i_t|| \leq \frac{L}{2}||w_t||,$$

which implies

$$\begin{aligned}\frac{d}{dt}V_\beta(y_t,w_t) &= -\sum_i D_{ii}\langle\hat{F}^i(y_t),\gamma^i_t\rangle - \sum_i D_{ii}\langle\hat{F}^i(y_t), f^i_\beta(D^{-1}Ay_t) - y^i_t\rangle - L||w_t||\\ &\leq -\sum_i D_{ii}\langle\hat{F}^i(y_t), f^i_\beta(D^{-1}Ay_t) - y^i_t\rangle - \frac{L}{2}||w_t||\\ &< -\sum_i D_{ii}\langle\hat{F}^i(y_t), f^i_\beta(D^{-1}Ay_t) - y^i_t\rangle.\end{aligned}$$

By lemma 9, we have for all $i\in V$ that

$$\langle\hat{F}^i(y_t), f^i_\beta(D^{-1}Ay_t) - y^i_t\rangle \geq 0,$$

which, together with the fact that $(\mathcal{F}^y_\beta,0)\subset\mathcal{K}\times B_{2\sqrt{N}}$ by item (ii), conclude the proof of item (iii).

(iv) First of all, one clearly has that $\mathcal{F}^y_\beta\subset\mathring{\Delta}^N_K$ by item (ii). Define the smooth function $\tilde{V}:\mathring{\Delta}^N_K\to\mathbb{R}_+$ as

$$\tilde{V}(y) = V_\beta(y,0).$$

Also, one has that the vector space $\mathcal{M}_{N\times K}(\mathbb{R})$ can be decomposed into $\mathcal{M}_{N\times K}(\mathbb{R}) = \mathcal{V}\oplus\mathcal{T}$, where $\mathcal{V} = \{\lambda\in\mathcal{M}_{N\times K}(\mathbb{R})\mid\lambda 1 = 0\}$, $\mathcal{T} = \{c1^T\mid c\in\mathbb{R}^N\}$, $\mathcal{V} = \mathcal{T}^\perp$ and $\mathcal{V}\cap\mathcal{T} = \{0\}$. Indeed, one has that for every $\lambda\in\mathcal{V}$ and $c\in\mathbb{R}^N$

$$\langle\lambda, c1^T\rangle = \langle\lambda 1, c\rangle = 0,$$

and in addition

$$dim(\mathcal{V}) + dim(\mathcal{T}) = N(K-1) + N = NK = dim(\mathcal{M}_{N\times K}(\mathbb{R})).$$

Now, we have by lemma 8 that for each $y \in \mathcal{F}_\beta^y$ there exists a vector $c(y) \in \mathbb{R}^N$ such that $\nabla_y V_\beta(y,0) = c(y)1^T \in \mathcal{T}$. Moreover, since $\mathring{\Delta}_K^N$ is an open set of an affine space with $\mathcal{V}$ as the associated vector space (see proof of lemma 6), we have that, since $c(y)1^T \in \mathcal{T} = \mathcal{V}^\perp$,

$$\nabla \tilde{V}(y) = Proj_{\mathcal{V}}\Big(\nabla_y V_\beta(y,0)\Big) = Proj_{\mathcal{V}}\Big(c(y)1^T\Big) = 0,$$

where $Proj_{\mathcal{V}} : \mathcal{M}_{N\times K}(\mathbb{R}) \to \mathcal{V}$ is the orthogonal projection of $\mathcal{M}_{N\times K}(\mathbb{R})$ onto $\mathcal{V}$.

Hence

$$\mathcal{F}_\beta^y \subset \{y \in \mathring{\Delta}_K^N \mid y \text{ is a critical point of } \tilde{V}\}$$

and the conclusion follows by Sard's lemma.

□

1.3.6 Proof of theorem 1

After proving in lemma 5 that the continuous time interpolation of (Y_t, W_t) is an Asymptotic Pseudotrajectory (APT) for the semiflow Φ generated by the limit ODE (1.11) (see definition 1), and by constructing a Lyapunov function for the set $(\mathcal{F}_\beta^y, 0)$ in lemma 10, we are ready to provide a rigorous proof of theorem 1 following the steps of the ODE method of subsubsection 1.3.1.2.

We first provide an auxiliary lemma regarding the sets $\mathcal{F}_\beta^x$ and $\mathcal{F}_\beta^y$:

Lemma 11. *Define the sets* $\mathcal{F}_\beta^y = \{y \in \Delta_K^N \mid y = f_\beta(D^{-1}Ay)\}$ *and* $\mathcal{F}_\beta^x = \{x \in \Delta_K^N \mid x = D^{-1}Af_\beta(x)\}$.

We have that $D^{-1}A\mathcal{F}_\beta^y = \mathcal{F}_\beta^x$, *where* $D^{-1}A\mathcal{F}_\beta^y = \bigcup_{y\in\mathcal{F}_\beta^y} D^{-1}Ay$.

Proof. Take $y \in \mathcal{F}_\beta^y$, then $x = D^{-1}Ay \in \Delta_K^N$ satisfies

$$x = D^{-1}Ay = D^{-1}Af_\beta(D^{-1}Ay) = D^{-1}Af_\beta(x),$$

hence $x \in \mathcal{F}_\beta^x$. This implies $D^{-1}A\mathcal{F}_\beta^y \subset \mathcal{F}_\beta^x$.

Take now $x \in \mathcal{F}_\beta^x$ and define $y = f_\beta(x) \in \Delta_K^N$. Then $x = D^{-1}Af_\beta(x) = D^{-1}Ay$ and

$$y = f_\beta(x) = f_\beta(D^{-1}Ay),$$

hence $y \in \mathcal{F}_\beta^y$. This implies $\mathcal{F}_\beta^x \subset D^{-1}A\mathcal{F}_\beta^y$ and concludes the proof. □

Proof of theorem 1: Let Φ be semiflow induced by the limit ODE (1.11), $(\overline{Y}, \overline{W})$ be the continuous time affine interpolation of (Y_t, W_t), both defined in lemma 2 and satisfying $P_t = D^{-1}AY_t + W_t$ for all $t \geq 0$, and also let $\mathcal{K} \times B_{2\sqrt{N}}$ be the invariant set for the semiflow Φ, defined by lemma 10.

First of all, since $(\overline{Y}, \overline{W})$ is an APT for Φ satisfying $(\overline{Y}(t), \overline{W}(t)) \in \mathcal{K} \times B_{2\sqrt{N}}$ for all $t \in \mathbb{R}_+$ by lemmas 5 and 10, we have that the semiflow Φ can be restricted without loss of generality to the set $\mathcal{K} \times B_{2\sqrt{N}}$, i.e., it can be defined as a function $\Phi_{|\mathcal{K}\times B_{2\sqrt{N}}} : \mathbb{R}_+ \times \left(\mathcal{K} \times B_{2\sqrt{N}}\right) \to \mathcal{K} \times B_{2\sqrt{N}}$.

Moreover, since $(\overline{Y}, \overline{W})$ is an APT for $\Phi_{|\mathcal{K}\times B_{2\sqrt{N}}}$, the limit set $L(Y,W) = \bigcap_{t\in\mathbb{R}_+} \overline{\bigcup_{s\geq t}(\overline{Y}(s), \overline{W}(s))}$ is internally chain transitive (see definition in [23]) by item (i) of theorem 5.7 in [23]. Finally, since V_β is a Lyapunov function for the compact and invariant set $(\mathcal{F}_\beta^y, 0)$ such that $V(\mathcal{F}_\beta^y, 0)$ has empty interior by lemma 10, we have by proposition 6.4 of [23] that $L(Y,W) \subset (\mathcal{F}_\beta^y, 0)$, which implies that $(Y_t, W_t) \to (\mathcal{F}_\beta^y, 0)$ almost surely when $t \to \infty$.

The uniform continuity of $D^{-1}A$ provides the almost sure convergence of $P_t = D^{-1}AY_t + W_t$ to the set $D^{-1}A\mathcal{F}_\beta^y$, which is equal to $\mathcal{F}_\beta^x$ by lemma 11. This concludes the proof.

□

1.4 Numerical examples

We perform in this section some numerical simulations of the proposed opinion dynamics algorithm, with the sole purpose of illustrating our claims. We use in the simulations 5 different undirected networks:

- The Zachary Karate Club (**ZKC**) network [314], with $N = 34$ people.
- The American College Football (**ACF**) teams in Division I during Fall 2000 regular season [116], with $N = 115$ teams.
- The social network of frequent associations between $N = 62$ bottlenose dolphins (**Dolphins**) over a period of seven years from 1994 to 2001 [214].
- An undirected randomly generated network following the Erdös Rényi model [96] (**Erdös Rényi**), with $N = 50$ nodes and an edge probability $p = 0.4$, i.e., edges are randomly sampled from independent Bernoulli random variables with parameter p.
- An undirected randomly generated network following the preferential attachment model of Barabási and Albert [21] (**Pref Attach**) with $N = 50$ nodes and 5 edges per new node, i.e., each new node creates 5 new edges in the network.

We simulate our opinion dynamics algorithm with the 5 networks above, using at each time the number of contents K equal to the number of nodes in the network, until a final time $T = 3000$ and with a softmax parameter[2] $\beta = 30$. Figures 1.2 and 1.3 depict the results of the simulations and provide a numerical validation of our main convergence result: theorem 1.

Figure 1.2 illustrates the convergence of the normalized scores P_t when $t \to \infty$ by plotting the evolution of their relative L^2 error at each 50 steps, i.e., it plots the function $rErr(t) = \frac{||P_{t+50} - P_t||}{||P_t||}$. The convergence of the function $rErr(t)$ to 0, represented by Figure 1.2, provides a numerical example of the almost sure convergence of the opinion dynamics algorithm (1.3), independently of the underlying network.

Figure 1.3 illustrates the convergence of the normalized scores P_t to the set $\mathcal{F}_\beta^x = \{x \in \Delta_K^N \mid x = D^{-1}Af_\beta(x)\}$ by plotting the L^2 error between the normalized scores P_t and the function $D^{-1}Af_\beta$ applied to the normalized scores, at each 50 steps, i.e., it plots the function $fErr(t) = ||P_t - D^{-1}Af_\beta(P_t)||$. The convergence of the function $fErr(t)$ to 0, represented by Figure 1.3, numerically confirms that indeed the normalized scores, under the opinion dynamics algorithm (1.3), converge almost surely to the set $\mathcal{F}_\beta^x = \{x \in \Delta_K^N \mid x = D^{-1}Af_\beta(x)\}$ when $t \to \infty$, independently of the underlying network.

1.5 Conclusion

We introduced a new opinion dynamics model which incorporates opinions about multiple contents and random broadcasts of information. The agents appreciation for each content is contained

2. We do not provide simulations with different softmax parameter values, since the main convergence theorem 1 remains valid for every softmax parameter value.

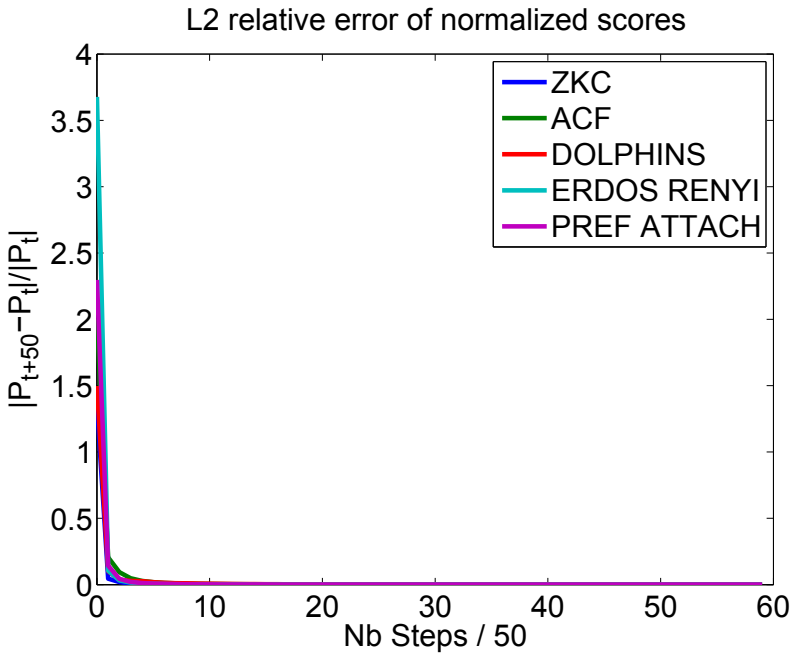

Figure 1.2: Plot of $rErr(t) = \frac{||P_{t+50} - P_t||}{||P_t||}$, the relative L^2 error of the normalized scores at each 50 steps, for each of the 5 networks: ZKC, ACF, DOLPHINS, ERDOS RENYI, PREF ATTACH.

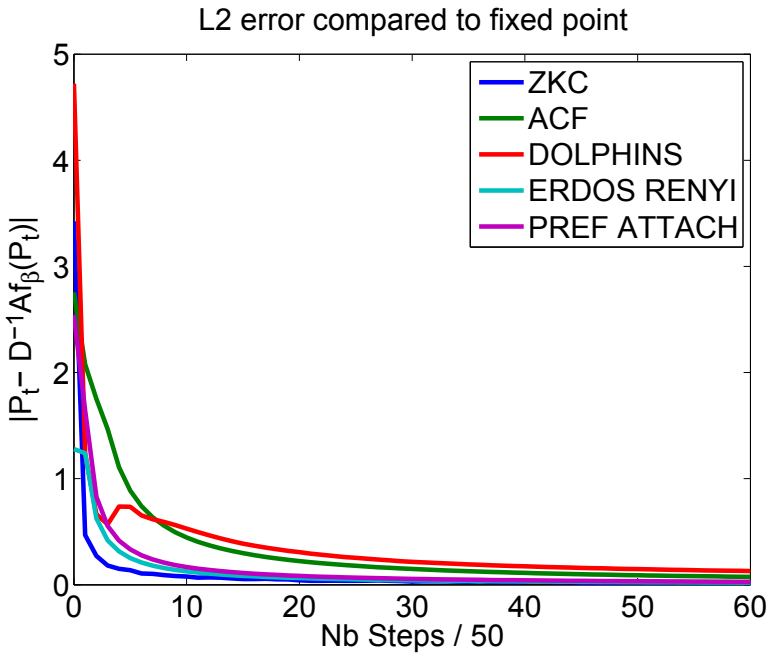

Figure 1.3: Plot of $fErr(t) = ||P_t - D^{-1}Af_\beta(P_t)||$, the L^2 error between the normalized scores and the function $D^{-1}Af_\beta$, at each 50 steps, for each of the 5 networks: ZKC, ACF, DOLPHINS, ERDOS RENYI, PREF ATTACH.

in an absolute score, and at each time step agents broadcast to their neighbors an opinion about a random content, which is chosen based on a softmax function of their relative scores. After this

broadcasting period, each agent interprets the received information and updates her scores in an additive fashion by weighting this new piece of information with her neighbors' influences over herself.

We showed that, when agents cannot influence themselves and influence each other equally, i.e., the network graph is undirected and without self-loops, there exists a Lyapunov function that provides the almost sure convergence of the opinion dynamics algorithm.

When $\beta \ll 1$, the algorithm converges to a consensus on the uniform distribution over the contents, and when $\beta \gg 1$, numerical simulations show that the algorithm converges to a point that represents a division of the network in clusters where agents inside each cluster only broadcast contents they appreciate the most.

Moreover, our main convergence theorem provides the convergence of agents' normalized scores for social networks with any number of agents, broadcasting information about any number of contents. It implies that our model remains coherent when simulating the opinion dynamics of agents using real-life social networks with millions of nodes.

CHAPTER 2

Community Detection

"Every person is defined by the communities she belongs to."

— Orson Scott Card, *Speaker for the Dead*

2.1 Introduction

This chapter introduces a novel community detection algorithm that discovers the network communities using the limit state of the opinion dynamics algorithm studied in chapter 1. The algorithm relies on the same principle than random walk methods for community detection. A random process takes place in the network, such that its limit state allows the discovery of the network communities.

As such, our algorithm bears resemblance to community detection methods stemming from statistical mechanics - such as the Potts-based clustering model of [29] - which are able to detect the communities of the network using the local minima of a Hamiltonian function. Moreover, our method relies on an underlying opinion dynamics algorithm where at each time step nodes randomly sample from a softmax distribution; this exponential weighting of states from the softmax distribution can be found for example in the Potts clustering algorithm [29], the zero-temperature Hamiltonian system equivalent to the label-propagation model of Raghavan *et al.* [256], and many other methods.

As discussed in chapter 1, the softmax parameter β impacts the random sampling of the opinion dynamics algorithm: a small value of β corresponds to a uniform choice over the K contents, and as β grows, the sampling becomes more biased towards contents with larger scores. In sight of this dichotomy, our community detection algorithm uses a large value of β (*e.g.* $\beta = 50$, $\beta = 100$, $\beta = 250$) for the random sampling step, which represents a large bias for contents with the highest scores.

Through the mechanism of the underlying opinion dynamics algorithm, one may realize that the proposed community detection algorithm resembles Von Dongen's Markov Cluster Algorithm (MCL) [293], which takes advantage of the transition probability matrix of a suitable random walker in the network and consists on the iteration of two steps: a first step called expansion, in which the transition matrix is raised to an integer power, and a second step called inflation, consisting in raising each entry of the transition matrix to some real-valued power and renormalizing this new matrix to be again a transition matrix of a random walk. The latter step enhances the weights between pairs of nodes with large edge weights, which are likely to belong to the same community;

after a few iterations, the process normally delivers a stable matrix, which can be associated with communities of the original network.

The broadcasting step of the opinion dynamics algorithm, which is the foundation of our community detection algorithm, works in a similar fashion as the inflation step of MCL, increasing the importance given to a content that already possesses a higher score. The communities are thus created around these contents. The main difference between both algorithms is the way how the expansion and inflation steps are performed: in our algorithm, the expansion step is achieved by a multiplication of the network's adjacency matrix, whereas MCL uses at each step a different transition matrix for the random walker; in the inflation step, our algorithm uses the softmax function with parameter β, whereas MCL uses a renormalized power of the actual transition matrix.

Since the softmax function with a large β parameter - used during the broadcasting step of the underlying opinion dynamics algorithm - takes into account only the contents with highest scores, we perform the same reduction as the label propagation method [256] - which is a particular instance of MCL: for each node, we only keep a small subset of active contents. As consequence, the communities found by our method bears resemblance to those of the label propagation method, described in [283].

Our method

i) can be mathematically proven to converge, contrary to heuristic state-of-the-art methods,

ii) describes the structure of the communities found,

iii) can be executed in a distributed fashion without difficulty,

iv) presents a manageable complexity,

v) discovers overlapping communities without an increase of complexity, and

vi) allows directed networks to be studied under the same practical framework, although it still lacks a theoretical proof of convergence.

vii) Moreover, it can be performed in two ways: a parametric and a nonparametric one; the parametric way is faster and allows the choice of the maximum number of detected communities; the nonparametric way does not cap the maximum number of discovered communities, thus overcharging the underlying opinion dynamics algorithm, increasing its complexity; this choice allows the proposed algorithm to behave as a multiscale community detection algorithm.

The rest of this chapter is organized as follows. Section 2.2 describes the community detection algorithm and introduces our definition of communities. Section 2.3 discusses the fine-tuning of the algorithm's parameters, in order to achieve an optimal performance, and the complexity of the algorithm in question. Section 2.4 performs some numerical tests and comparisons with other community detection methods. And Section 2.5 concludes the chapter.

2.2 The community detection algorithm and definition of communities

2.2.1 Notations

We use the notations of chapter 1, as described in subsection 1.2.1. Moreover, we denote for the matrix M its (i,k) entry as $M_{i,k}$ or $M^{i,k}$, the L^∞-norm $|M|_\infty = \max_{i,k} |M_{ik}|$, the spectral radius

$sp(M) = \sup\{|\lambda| \mid det(M - \lambda\mathbb{I}) = 0\}$, and $v : \mathcal{M}_{N\times K}(\mathbb{R}) \to \mathbb{R}^{NK}$ the vectorization operation of matrices.

We define for any real numbers x and y, $\lfloor x \rfloor$ as the floor of x, $\lceil x \rceil$ as the ceiling of x, $x \wedge y = \min\{x, y\}$ as the minimum between x and y and $x \vee y = \max\{x, y\}$ as the maximum between x and y and we denote for every set $\mathcal{A}$ its cardinality as $|\mathcal{A}|$.

For two positive real functions f, g, we also denote $f(x) \sim \mathcal{O}(g(x))$ (or simply $f \sim \mathcal{O}(g)$) if and only if there exists two constants $M, x_0 \geq 0$ such that $f(x) \leq Mg(x)$ for all $x \geq x_0$.

2.2.2 The community detection algorithm

We now begin the main focus of this chapter, the derivation of our community detection algorithm. As already mentioned, the foundation of our community detection algorithm is the opinion dynamics model of chapter 1. The community detection algorithm has the following steps, which are written in a more concise form in algorithm 1:

1 - Choose the parameters of the community detection algorithm:

- 1.1 - T: number of steps of the opinion dynamics algorithm.
- 1.2 - $\beta \gg 1$: softmax parameter for the random sampling under the distribution in Eqn. (1.1).
- 1.3 - K: number of contents of the opinion dynamics algorithm.
- 1.4 - X_0: initial condition of the opinion dynamics algorithm.

2 - For each time step $0 \leq t \leq T - 1$:

- 2.1 - Create an auxiliary $N \times K$ matrix Aux_{t+1} and initialize it to 0.
- 2.2 - Get the normalized score matrix P_t from the nodes score matrix X_t as

$$P_t^{i,k} = \frac{X_t^{i,k}}{\sum_{k'} X_t^{i,k'}}.$$

- 2.3 - For each node $i \in \{1, 2, \cdots, N\}$:
  - 2.3.1 - Sample the content to be broadcasted by node i at time $t + 1$, denoted by I_{t+1}^i, following the random law given by Eqn. (1.1), as

$$\mathbb{P}(I_{t+1}^i = k) = f_\beta^{i,k}(P_t).$$

  - 2.3.2 - Broadcast the content I_{t+1}^i to the neighbors of node i, i.e., for each node $j \sim i$:
    - 2.3.2.1 - Increment the entry $(j, \mathcal{I}_{t+1}^i)$ of the auxiliary matrix Aux_{t+1} with the weight of the edge between nodes i and j, as

$$Aux_{t+1}^{j,I_{t+1}^i} = Aux_{t+1}^{j,I_{t+1}^i} + A_{j,i}.$$

- 2.4 - Update the nodes scores by adding the auxiliary matrix Aux_{t+1} to the ancient scores as

$$X_{t+1} = X_t + Aux_{t+1}.$$

3 - Retrieve the communities $(c_k)_{k\in\{1,2,\cdots,K\}}$ of G from the final normalized scores P_T as

$$c_k = \{i \in V \mid P_T^{i,k} \geq \max_l P_T^{i,l} - \delta(\beta)\}, \tag{2.1}$$

where $\delta(\beta) \ll 1$. As a rule of thumb, one may use $\delta(\beta) = 1/\sqrt{\beta}$ or $\delta(\beta) = 1/\beta$.

Algorithm 1 - Opinion Dynamics Clustering

Input: Graph $G = (V, E)$
Choose T the number of steps, the softmax parameter $\beta \gg 1$, the number of contents K and the initial condition X_0.
Create an auxiliary $N \times K$ matrix Aux.
for each time step $0 \leq t \leq T-1$ **do**
    Initialize $Aux \leftarrow 0$.
    Get normalized scores P_t following $P_t^{i,k} \leftarrow \frac{X_t^{i,k}}{\sum_{k'} X_t^{i,k'}}$.
    for each node $i \in \{1, 2, \cdots, N\}$ **do**
        Sample the content I_{t+1}^i, following $\mathbb{P}(I_{t+1}^i = k) = f_\beta^{i,k}(P_t)$.
        for each neighbor node $j \sim i$ **do**
            Increment entry (j, I_{t+1}^i) of auxiliary matrix Aux as

$$Aux^{j,I_{t+1}^i} \leftarrow Aux^{j,I_{t+1}^i} + A_{j,i}.$$

        end for
    end for
    Update nodes scores following

$$X_{t+1} \leftarrow X_t + Aux$$

end for
Retrieve the communities $\mathcal{C} = (c_k)_{k \leq K}$ following Eqn. (2.1) as

$$c_k = \{i \in V \mid P_T^{i,k} \geq \max_l P_T^{i,l} - \delta(\beta)\}.$$

Output: Communities $\mathcal{C} = (c_k)_{k \leq K}$.

2.2.3 Definition of communities

Let us assume, for simplicity, that nodes inside a community have only one preferred content (which is the same for every node in the community) and let us associate the communities with the preferred content of nodes belonging to them, denoting a community by c_k if the preferred content is $k \in \{1, 2, \cdots, K\}$. In sight of algorithm 1, when $\beta \gg 1$, we expect that outside c_k, there exists a smaller number of nodes that possess their highest scores relative to content k, and that the nodes inside c_k possess more edges inside c_k than edges flowing outside c_k, since otherwise they would receive more information about a different content and they would possibly end up with a higher score for this content when compared to content k. This motivates the following definition of communities:

Let $\mathcal{P}(N) = 2^V$ be the power set of $V = \{1, 2, \cdots, N\}$ and $\mathcal{C} = (c_k)_{k \in \{1,2,\cdots,K\}} \subset \mathcal{P}(N)$ be a set of sets. We can define for each $i \in V$ the probabilities $p_{i,k}$ as

$$p_{i,k} = \begin{cases} \frac{1}{|\{k' \mid i \in c_{k'}\}|} \text{ if } i \in c_k \\ 0 \text{ if } i \notin c_k, \end{cases} \tag{2.2}$$

i.e., for each $i \in V$ we can define probabilities $p_{i,k}$ that are uniform on the sets c_k that contain i. These probabilities represent the sets of nodes: if a node i has $p_{i,k} > 0$ it means that it belongs to the community c_k.

Definition 3. *The graph G is divided in communities $\mathcal{C} = (c_k)_{k \leq K}$ if*

- $\bigcup_k c_k = V$.
- *for each $c_k \neq \emptyset$ we have the following: for every $i \in c_k$,*

$$\sum_{j \in c_k} A_{ij} p_{j,k} \geq \sum_{j \in c_{k'}} A_{ij} p_{j,k'}, \ \forall k' \neq k. \tag{2.3}$$

In other words, a community c_k is a subgroup of nodes of G such that for each node i in c_k, the weighted sum of the probabilities of the neighbors of i in c_k is larger than or equal to the weighted sum of the probabilities of the neighbors of i in any other community, compared in a pairwise fashion. This definition is different from the usual definition of communities found in the literature [255], but it also reproduces the intuition that there are clusters of nodes more connected within themselves than with other clusters.

Remark: If for every node $i \in V$, there exists only one nonzero entry for $(p_{i,k})_{k \leq K}$, then node i belongs to a single community and condition (2.3) becomes: for every $i \in c_k$,

$$\sum_{j \in c_k} A_{ij} \geq \sum_{j \in c_{k'}} A_{ij}, \ \forall k' \neq k,$$

which is exactly the same property of the communities found by the label propagation method [256], as explained in [283].

This definition of communities is intimately related to a certain set $\mathcal{F}_\infty \subset \Delta_K^N$, as explained by the next proposition:

Proposition 1. *Let $f_\infty : \mathcal{M}_{N \times K}(\mathbb{R}) \to \Delta_K^N$ be defined as*

$$f_\infty^{i,k}(p) = \frac{\mathbb{I}_{\{k \in \mathcal{M}_i(p)\}}}{|\mathcal{M}_i(p)|}, \tag{2.4}$$

with $\mathcal{M}_i(p) = \{k \mid p^{i,k} = \max_{k'} p^{ik'}\}$ *the set with the highest indices of* $p^i = (p^{i,1}, \cdots, p^{i,K})$*, i.e., the set with indices of maximum values for node* i *with respect to the matrix* p*, and define as well the set*

$$\mathcal{F}_\infty = \{x \in \Delta_K^N \mid x = D^{-1}Af_\infty(x)\}.$$

Let $x \in \mathcal{F}_\infty$ *and define the sets*

$$c_k^x = \{i \in V \mid f_\infty^{i,k}(x) > 0\}. \tag{2.5}$$

Then $\mathcal{C}^x = (c_k^x)_{k \le K}$ *is a division of* G *in communities.*

Proof. Since $\sum_k f_\infty^{i,k}(x) = 1$ for every $i \in V$, we have that there exists at least one nonzero entry in $(f_\infty^{i,k}(x))_{k \le K}$, and thus every node i belongs to some community. This trivially implies $\bigcup_k c_k^x = V$.

We also have, by the definition of $\mathcal{C}^x$, that $f_\infty^{i,k}(x) > 0 \Leftrightarrow i \in c_k^x$, which implies

$$i \in c_k^x \Leftrightarrow f_\infty^{ik}(x) > 0 \Leftrightarrow x^{i,k} = \max_l x^{i,l} \Leftrightarrow k \in \mathcal{M}_i(x)$$

and

$$f_\infty^{i,k}(x) = \frac{\mathbb{I}_{\{k \in \mathcal{M}_i(x)\}}}{|\{k' \mid k' \in \mathcal{M}_i(x)\}|} = \frac{\mathbb{I}_{\{i \in c_k^x\}}}{|\{k' \mid i \in c_{k'}^x\}|} = p_{i,k},$$

by the definition of $p_{i,k}$ in Eqn. (2.2).

Let $i \in c_k^x$, then $x^{i,k} \ge x^{i,k'}$ for all $k' \ne k$, and

$$\sum_j A_{ij}p_{j,k} = \sum_j A_{ij}f_\infty^{j,k}(x) = D_{ii}(D^{-1}Af_\infty(x))_{i,k} = D_{ii}x^{i,k} \ge D_{ii}x^{i,k'} = \sum_j A_{ij}p_{j,k'},\ \forall k' \ne k.$$

Since $p_{j,k} \ne 0 \Leftrightarrow j \in c_k^x$, this shows that for each k and for all $i \in c_k^x$ we have

$$\sum_{j \in c_k^x} A_{ij}p_{j,k} \ge \sum_{j \in c_{k'}^x} A_{ij}p_{j,k'},\ \forall k' \ne k$$

and $\mathcal{C}^x$ is indeed a division of G in communities. □

Moreover, we also have the following lemma about the minimum size of a community:

Lemma 12. *Let* $(c_k)_{k \le K}$ *be a division of* G *in communities. If for every* $i \in V$ *we have that* $A_{ii} = 0$*, then* $|c_k| \ge 2$ *for every nonempty community* c_k*.*

Proof. Let us assume without loss of generality that node 1 belongs to community c_1 with $|c_1| = 1$, i.e., $c_1 = \{1\}$, and let us prove the result by contradiction.

By assumption D.1(i), there exists a node $i \in V$ such that $A_{1i} > 0$, which belongs to a community $c_{k_i} \ne c_1$. Thus, by Eqn. (2.3) in the definition of communities, we have that (since $A_{11} = 0$)

$$0 = \sum_{j \in c_1} A_{1j}p_{j,1} \ge \sum_{j \in c_{k_i}} A_{1j}p_{j,k_i} \ge A_{1i}p_{i,k_i},$$

which implies $p_{i,k_i} = 0$ since $A_{1i} > 0$. This is a contradiction since $i \in c_{k_i} \Leftrightarrow p_{i,k_i} > 0$ by Eqn. (2.2). □

2.2.4 Discussion

As already mentioned in chapter 1, the opinion dynamics algorithm (1.3) behaves quite differently when varying the softmax parameter β: if $\beta \ll 1$, nodes sample the content to be broadcasted at each time step almost uniformly, whereas the broadcasting step with $\beta \gg 1$ is biased towards the contents possessing the highest score.

As consequence, when $\beta \gg 1$, we expect that if a node i has a higher number of neighbors that prefer the same content k, the limit normalized scores of node i should also bear a higher value for content k, which implies that a limit point for P_t should somehow "cluster" the network G in communities where nodes inside the same community have the highest scores over the same contents. This is the main intuition behind the community detection algorithm 1.

Despite the simplicity of the community detection algorithm 1, there are two important points that need to be discussed:

1. The numerical experiments performed in section 1.4 of chapter 1 suggest that the normalized scores P_T converge almost surely, when $T \to \infty$, to an element of the limit set $\mathcal{F}_\beta^x$ (see figure 1.2). Moreover, Peter Tino already proved in [285, 286] that for the vectorial case (where the adjacency matrix A is the identity matrix $\mathbb{I}$ and nodes do not affect each other), the set $\mathcal{F}_\beta^x$ converges to the set $\mathcal{F}_\infty$ when $\beta \to \infty$, in the sense that $\sup_{x' \in \mathcal{F}_\beta^x} d(x', \mathcal{F}_\infty) \to 0$ when $\beta \to \infty$, i.e., for every sequence $x^\beta \in \mathcal{F}_\beta^x$ we have that $\lim_{\beta \to \infty} x^\beta \in \mathcal{F}_\infty$ (since $\mathcal{F}_\infty$ is finite).

   This item provides the motivation of why the community detection algorithm 1 discovers the communities of G by using Eqn. (2.1): it relates a point in $\mathcal{F}_\beta^x$ to the discovered communities of G, which, since $\mathcal{F}_\beta^x \sim \mathcal{F}_\infty$ when $\beta \gg 1$, satisfy proposition 1.

2. The discovery of the network communities is performed by Eqn. (2.1), using the final normalized scores P_T provided by the underlying opinion dynamics algorithm (1.3). The choice of retrieving the communities using this equation is by no means the only way of retrieving the network communities. This choice is based on proposition 1, which describes the set $\mathcal{F}_\infty$ (which is close to the limit set $\mathcal{F}_\beta^x$ when $\beta \gg 1$) as a subset of the communities of G, following definition 3.

   Thus, the division of $G = (V, E)$ into communities $\mathcal{C} = (c_k)_{k \in \{1,2,\cdots,K\}}$ is associated with node probabilities $(p^i)_{i \in V}$ such that $p^{i,k}$ is the probability that node i belongs to community c_k (see Eqn. (2.2)), in such a way that each community has the same impact on node i if node i belongs to more than one community. This association of probabilities p to nodes given communities c_k provides, by proposition 1, a theoretical guarantee that each discovered community possesses "sufficient" mass, in the sense that for each discovery community c_k, the weighted sum of edges of nodes in c_k - weighted by the nodes probabilities p in Eqn. (2.3) - is greater than or equal to the weighted sum of edges flowing to another community $c_{k'}$, when comparing c_k against every other community $c_{k'} \neq c_k$ in a pairwise fashion.

   Again, this association of nodes probabilities p to communities $\mathcal{C} = (c_k)_{k \in \{1,2,\cdots,K\}}$, as illustrated by Eqn. (2.2), is by no means unique and one could provide different associations. For example, one could associate for each node i the probabilities $P_T^i = (P_T^{i,1}, \cdots, P_T^{i,K})$ such that for every content k giving rise to a community c_k we have that $P_T^{i,k}$ is the probability that node i belongs to the community c_k. These probabilities are of course different from the probabilities p defined by Eqn. (2.2), which can be retrieved from P_T by means of Eqn. (2.1).

   The retrieval of p (defined by Eqn. (2.2) and calculated from the discovered communities using Eqn. (2.1)) from P_T can be hence seen as a soft thresholding of P_T in the sense that all

nonzero probabilities P_T that are not above some threshold (depending itself on the maximum value of P_T^i for each node i) are reduced to 0 in order to generate p.

That being said, one argument in favor of our choice of nodes probabilities p given communities $\mathcal{C}$, as illustrated by Eqn. (2.2), is the theoretical guarantee of proposition 1, which is no longer true for probabilities different than p.

Remark: We discover the communities with Eqn. (2.1) in order to accommodate overlapping communities, in which one node may belong to more than one community. However, if one is not interested in overlapping communities, she can find the communities c_k given by Eqn. (2.1) as

$$c_k = \{i \in V \mid P_T^{i,k} = \max_l P_T^{i,l}\}.$$

2.3 Choice of parameters and complexity

One of the strong points of the proposed opinion-dynamics-based community detection algorithm is its simplicity. This community detection algorithm simply updates scores (which can be performed in a distributed fashion) using Eqn. (1.2) and retrieves communities using the simple rule of Eqn. (2.1). However, to reach maximum performance, one may need to tune the parameters of the algorithm: the initial condition X_0, the softmax parameter β, the number of steps T of the opinion dynamics algorithm and the number K of different contents in the opinion dynamics algorithm.

This section focuses on how to choose the parameters and when to stop the opinion dynamics algorithm (1.2) to retrieve the discovered communities, in order to decrease the complexity of the procedure and increase the odds of finding nontrivial communities.

2.3.1 Initial condition and number of contents

One can clearly see that a division of G in communities can always be achieved by $c_k = V$ and $c_{k'} = \emptyset$ for all $k' \neq k$, which is associated with an element $x \in \mathcal{F}_\infty$ such that $x^{i,l} = \mathbb{I}_{\{l=k\}}$ for some fixed $k \in \{1, \cdots, K\}$. These communities are trivial ones, and one must make sure that our opinion dynamics algorithm does not converge to one of them when $t \to \infty$. Since we do not have control of the limit point of the opinion dynamics algorithm, a suitable initial condition must be chosen in order to assure the convergence to nontrivial communities.

We denote by $A \propto B$ if matrix A is proportional to matrix B, i.e., $A = \gamma B$ with $\gamma > 0$. A "good" generic initial condition X_0 that has been numerically tested that converges to a nontrivial division of the graph is $X_0 \propto A$, with $K = N$, i.e., the number of contents equals the number of nodes in V. This is a very logical choice: since we have no information on the number of communities that our algorithm will find, we take K as big as possible to accommodate every possibility, hence we choose $K = N$.

The choice $X_0 \propto A$ gives $P_0 = D^{-1}A$ and highlights a similarity with the label propagation algorithm [256]: when $K = N$, each content can be associated with a community and we can say that, at first, each node i belongs to its own community, labeled with an abuse of notation i. Thus, at the first iteration, nodes choose one content to broadcast randomly to their neighbors, using the softmax function; when the graph G is unweighted, i.e., $A_{ij} \in \{0, 1\}$, this choice is made uniformly. At each iteration, nodes start to change communities, being influenced by their neighbors' broadcasts. At the end, nodes belong to the communities corresponding to the content they received the most. When communities do not overlap, we retrieve thus the result of [283].

This choice of initial condition gives systematically a nontrivial division of communities - a fact sustained by an extensive number of simulations. This happens because, at first, nodes broadcast only information about contents that represent their neighbors (recall that $P_0^{i,k} = \frac{A_{ik}}{D_{ii}}$), thus they will continue broadcasting contents relative to their neighbors (which are associated with a community since $K = N$), and the system converges to a configuration where the normalized scores concentrate on the neighbors possessing the highest degree. Hence, these high-degree nodes transmit their communities to their neighbors, which in their turn transmit to their own neighbors, and gradually cluster the network, as desired.

On the other hand, one may want to cap the maximum number of communities to be found, let us say, to $K < N$. This limitation on the number of communities may stem from two main reasons: first, the complexity of the algorithm increases with K, since at each iteration we need to sample N random variables from a K-dimensional vector. Second, one may want to retrieve a smaller number of communities with a larger number of nodes inside each community, thus the capping of the maximum number of communities forces the nodes to redistribute themselves among a smaller number of communities and increase their sizes (this phenomenon is represents the *multiscale* character of community detection, which is studied at length in [103, 287]).

We can retrieve a capped similar initial condition as follows: enumerate the nodes from 1 to N and divide them into K blocks of size $\lceil N/K \rceil$ (note that the last block will have a size smaller than or equal to $\lfloor N/K \rfloor$). Then we define, for $k \in \{1, 2, \cdots, K\}$, the initial condition X_0 as

$$X_0^{i,k} \propto \sum_{j=(k-1)\times\lceil N/K\rceil+1}^{(k\times\lceil N/K\rceil)\wedge N} A_{ij}.$$

Hence $X_0^{i,k}$ represents the proportion of neighbors of node i in block k, for $k \in \{1, 2, \cdots, K\}$. The case $K = N$ happens when we have only one node per block.

Remark: Using lemma 12 we can choose $K \leq \lceil N/2 \rceil$ and X_0 accordingly, since every community must have at least two nodes when the network is without self loops.

Remark: If one can estimate in advance the maximum number of communities in G, then applying the beforementionned choice of K and X_0 drastically decreases the complexity of the algorithm. Also, if one already knows some of the communities, she can bias the initial condition in order to "direct" the opinion dynamics algorithm to converge faster to the desired communities.

2.3.2 Running time T and softmax parameter β

The choice of the number of steps T for the opinion dynamics algorithm to achieve convergence is of utmost importance, since it is the major contributor for the complexity of the community detection algorithm, as well as the softmax parameter β, since it is responsible for the approximation of f_∞ (defined by Eqn. (2.4)) by f_β when $\beta \gg 1$. Clearly, both parameters depend on the number of nodes N, the number of contents K and on the structure of the graph G.

Theoretical bounds for the running time T using the convergence of the stochastic approximation algorithm (1.9) could be obtained from laws of iterated logarithm [177] or central limit theorems [248]. However, these bounds are not satisfactory for three main reasons: first, one needs to compute the asymptotic covariance of the martingale differences ζ_{t+1} in Eqn. (1.9), second, the results stemming from laws of iterated logarithms and central limit theorems are asymptotic and do not provide an analytic lower bound, and third, these generic bounds are most of the time conservative and do not exploit the full structure of the model.

For example, let us take $P_T \to P_\beta \in \mathcal{F}_\beta^x$ almost surely when $T \to \infty$. By the almost sure central limit theorem 1 of [248] that, conditional to the event $P_T \to P_\beta$, we have

$$\sqrt{T}(v(P_T) - v(P_\beta)) \xrightarrow{d} \mathcal{N}(0, \Gamma),$$

where Γ is the $NK \times NK$ asymptotic covariance matrix taking into consideration the covariance of $v(\zeta_{T+1})$ and the derivative of f_β at the limit point P_β.

Assuming that errors smaller than $\frac{1}{N}$ do not affect the communities found by our algorithm (because these communities are associated with points $P_\beta \sim x \in \mathcal{F}_\infty$ when β is large enough), we must control the probability that $||v(P_T) - v(P_\beta)||$ is greater than $\frac{1}{N}$. Since $y^T\Gamma^{-1}y \geq \frac{||y||^2}{sp(\Gamma)}$ for all vectors $y \in \mathbb{R}^{NK}$ by the partial ordering on the symmetric positive-definite matrices, we have by the almost sure central limit theorem that

$$\begin{aligned}
\mathbb{P}(||v(P_T) - v(P_\beta)|| \geq \frac{1}{N}) &= \mathbb{P}(\frac{T}{sp(\Gamma)}||v(P_T) - v(P_\beta)||^2 \geq \frac{T}{sp(\Gamma)N^2}) \\
&\leq \mathbb{P}((v(P_T) - v(P_\beta))^T(\frac{\Gamma}{T})^{-1}(v(P_T) - v(P_\beta)) \geq \frac{T}{sp(\Gamma)N^2}) \\
&= \mathbb{P}(\chi^2_{NK} \geq \frac{T}{sp(\Gamma)N^2}),
\end{aligned}$$

where χ^2_{NK} is a Chi-Squared random variable with NK degrees of freedom. Since $\frac{\chi^2_{NK} - NK}{\sqrt{2NK}} \xrightarrow{d} \mathcal{N}(0,1)$ when $NK \to \infty$, we have that for a 5% probability of $||v(P_T) - v(P_\beta)||$ being greater than $\frac{1}{N}$, we must have

$$\frac{T}{sp(\Gamma)N^2} \in [NK - 2\sqrt{2NK}, NK + 2\sqrt{2NK}] \Rightarrow T \sim \mathcal{O}(N^3 K sp(\Gamma)).$$

One has, at least intuitively, that $sp(\Gamma)$ decreases when $\beta \nearrow \infty$ because Γ takes into consideration the quality of the approximation of f_∞ by the softmax function f_β (or in other terms, the convergence of the set $\mathcal{F}_\beta^x$ towards the set $\mathcal{F}_\infty$), which means that increasing β reduces the asymptotic variance of Γ, and by consequence, the running time T.

Two important details must be discussed here:

- Although we have that increasing the softmax parameter β decreases the running time T for the opinion dynamics algorithm, it has a lower bound by the central limit theorem, i.e., the almost sure central limit theorem assures the convergence of $\sqrt{T}(v(P_T) - v(P_\beta))$ to a normal distribution of covariance matrix Γ only when $T \to \infty$, thus we cannot increase β indefinitely and expect to reduce the running time T.

  Thus, we still cannot provide a theoretical lower bound for the running time T, which presents itself as quite challenging. With this lack of theoretical results, we used in our experiments in section 2.4 bounds on T of the form

  $$T \sim \mathcal{O}(\log N) \quad \text{or} \quad T \sim \mathcal{O}(\sqrt{N}),$$

  nevertheless a more detailed study should be conducted.

- On the other hand, if care is not taken in the choice of β, one may have very large values of $e^{\beta P_t^{i,k}}$ that may be larger than the maximum boundaries for floating numbers in the computer

(as an example, $e^{100} \sim 2.6 \times 10^{43}$), which would provoke numerical errors when sampling the random variables $\mathcal{I}_{t+1}^i$ from the softmax function f_β in Eqn. (1.1), invalidating the sampling mechanism and as a result invalidating the community detection algorithm entirely.

Thus, the choice of β must lie in an optimal interval in order for the community detection algorithm to achieve a good performance. However, finding this optimal interval for the softmax parameter β is still a challenging task. We used for our simulations in section 2.4 a softmax parameter β varying from $50, 100, 200, 250$, which again needs a more detailed and structured study.

2.3.3 Complexity of the community detection algorithm

We calculate now the complexity of our community detection algorithm. At each step $t \in \{1, 2, \cdots, T\}$ of our opinion dynamics algorithm (1.2), each node $i \in V$ samples a content $k \in \{1, 2, \cdots, K\}$ to broadcast to his $d_i = \sum_j \mathbb{I}_{\{A_{ji}>0\}}$ neighbors.

The sampling operation from the softmax function is of order $\mathcal{O}(K)$, and broadcasting it is of order $\mathcal{O}(d_i)$. Hence, at the end of broadcasting at a single time step t, we have a complexity of

$$\sum_i \mathcal{O}\big(K + d_i\big) = \mathcal{O}(|V| \times K + |E|) = \mathcal{O}(NK + |E|).$$

After storing every content to be updated from the broadcasting step t, we must update the scores X_{t+1}. This operation is of order $\mathcal{O}(NK)$ since we are simply summing up two matrices of size $N \times K$.

At the end of the T time steps of our opinion dynamics algorithm, we have the complexity

$$\begin{aligned} &\mathcal{O}(T \times (NK + |E|)) \quad \text{from the broadcasting} \\ &\mathcal{O}(T \times NK) \quad \text{from the updating,} \end{aligned}$$

which is of complexity

$$\mathcal{O}(T \times (NK + |E|)).$$

Finally, using proposition 1, we retrieve the communities using Eqn. (2.1). We must thus calculate the maximum of N vectors with K coordinates, which is of complexity $\mathcal{O}(NK)$.

In summary, after the T broadcasting and updating steps and with the retrieval of communities, our algorithm has total complexity

$$\mathcal{O}\Big(T \times \big(NK + |E|\big) + NK\Big) = \mathcal{O}\Big(T \times \big(NK + |E|\big)\Big).$$

Remark: If one does not have enough memory to store the totality of the updates at each step, one can update the scores individually at each broadcast, randomly choosing a node i to perform the broadcast at each time step t.

Remark: One can also readily check that the complexity of the algorithm is much higher for the nonparametric version, since in this case $K = N$ or $K = N/2$ by lemma 12. For the parametric version, one can choose a much lower value for K, which decreases significantly the complexity of the algorithm.

2.3.4 Speeding up the algorithm

There are several ways of speeding up this community detection algorithm, all of them based on some heuristics. For example:

- Suppose that we have a total number of time steps T_K for the opinion dynamics algorithm with K contents. Starting our algorithm with $X_0 \propto A$ and running it until some time $t_1 \ll T_K$, we can check whether some of the normalized scores $(P_t^{i,k})_{i \in V}$ are close to 0 for certain contents k. If it is indeed true, it means that these contents will always be close to 0 during the algorithm, thus we can eliminate their entire columns from P_{t_1} and rerun the algorithm with a smaller number of contents. This reduces the complexity of the algorithm and allows it to run with a number of contents closer to the actual number of communities to be found.

- Use the true function f_∞ instead of f_β for time steps $t \gg 1$, since the normalized scores P_t will be already converging to the limit point that generates the communities. Sampling from f_β is costlier than from f_∞, but it is absolutely necessary in the earlier stages of the opinion dynamics algorithm where the initial condition is far away from the limit points.

2.4 Numerical examples

This section is dedicated to numerical examples of our community detection algorithm. We perform a comparison between our algorithm and some benchmark algorithms found in the literature, for six undirected networks:

- Undirected Zachary Karate Club (**ZKC**) network [314], with $N = 34$ people. See figure 2.1.

- Undirected American College Football (**ACF**) teams in Division I during Fall 2000 regular season [116], with $N = 115$ teams. See figure 2.2.

- Undirected social network of frequent associations between $N = 62$ bottlenose dolphins (**Dolphins**) over a period of seven years from 1994 to 2001 [214]. See figure 2.3.

- Undirected network composed of 10 friend lists (ego networks) from Facebook (**Facebook-ego**) [223], with $N = 4,039$ users. See figure 2.4.

- Undirected general relativity and quantum cosmology collaboration network in ArXiv (**GRQC-ArXiv**) with $N = 5,241$ researchers, from January 1993 to April 2003 [196]. See figure 2.5.

- Undirected contact network between the $N = 15,088$ users in Youtube (**Youtube**), crawled in December 2008 [278]. See figure 2.6.

We compare our community detection method with different algorithms: the label propagation method [256], the Louvain method [35], greedy implementations of the modularity [236] and the statistical mechanics Hamiltonian of Reichardt and Bornholdt [260], and fast implementations of them based on the algorithm of Le Martelot and Hankin [191] using global criteria.

We denote by:

- **OD** our opinion-dynamics-based community detection algorithm.

- **LP** the label propagation method of [256].

- **L1**, **L2**, **L3** different clustering divisions given by the Louvain method [35].

- **FM** an implementation of the fast algorithm using global criteria of [191] with as optimization criterion the modularity function of [236].
- **FN** a greedy hierarchical method to maximize the modularity function, proposed by Newman in [233]. It works as follows: first one defines every node as its own community, and at each step one computes for each pair of communities the difference in modularity when merging them, proceeding in the direction of the highest gain in modularity (stopping the algorithm if there is none).
- **SRB1**, **SRB2**, **SRB3** different implementations of the fast algorithm using global criteria of [191] with as optimization criterion the Reichardt and Bornholdt Hamiltonian function [260] with scale parameters $\gamma = 0.8, 0.7, 0.6$, respectively.
- **R** a greedy hierarchical method (defined before for the method FN) to maximize the Hamiltonian function of Reichardt and Bornholdt [260], with a scale parameter $\gamma = 0.8$.

We compute for each community detection algorithm comparative measures on the beforementioned six undirected networks. The comparative measures are

- Modularity (**mod**).
- Number of communities (**nbC**).
- Average community density (**den**), i.e., average ratio between number of edges inside communities and number of pairs inside communities.
- Average community embeddedness (**emb**), i.e., average ratio between number of edges from inside communities and the total number of edges of the communities.
- Average proportion of the communities to the total number of nodes (**siz**).
- Normalized mutual information [188] between the communities found by other methods and those found by our method (**mut**), i.e., a number between 0 and 1 that checks the similarity of the communities.
- *And when available,* the normalized mutual information between the communities found and the ground truth communities (**miTrue**), which measures how well the algorithms were in discovering the true communities.

2.4.1 ZKC and ACF

We start by comparing our algorithm with benchmarks algorithms over the ZKC (table 2.1) and ACF (table 2.2) networks. These networks are small but important, due to the fact that they are the only ones whose information about their true communities is available.

One can see in table 2.1 that our algorithm (and the label propagation method, who found the same communities as we did) was by far the best in reconstructing the true communities of the ZKC. It is important to notice that 84% is relatively big, because we only have 34 nodes, missing one node reduces drastically the normalized mutual information (in fact we miss only 1 node). Our algorithm also finds the communities with the highest average size, lowest average density and highest average embeddedness (together with the label propagation method and the greedy Reichardt and Bornholdt statistical mechanics method). This suggests that, in the ZKC example,

	OD	LP	L1	L2	L3	FM	FN	SRB1	SRB2	SRB3	R
mod	0.37	0.37	0.37	0.42	0.42	0.42	0.39	0.48	0.52	0.57	0.47
den	0.25	0.25	0.64	0.44	0.44	0.45	0.35	0.40	0.40	0.25	0.25
siz	0.50	0.50	0.17	0.25	0.25	0.25	0.33	0.33	0.33	0.50	0.50
mut	1.00	1.00	0.52	0.59	0.59	0.69	0.69	0.70	0.70	0.84	0.84
nbC	2	2	6	4	4	4	3	3	3	2	2
emb	0.77	0.77	0.38	0.56	0.56	0.56	0.59	0.66	0.66	0.77	0.77
miTrue	0.84	0.84	0.44	0.49	0.49	0.59	0.56	0.57	0.57	0.68	0.68

Table 2.1: Comparative table ZKC

	OD	LP	L1	L2	L3	FM	FN	SRB1	SRB2	SRB3	R
mod	0.58	0.60	0.60	0.60	0.60	0.60	0.55	0.63	0.65	0.67	0.58
den	0.88	0.82	0.86	0.76	0.76	0.76	0.48	0.60	0.60	0.54	0.52
siz	0.08	0.09	0.08	0.10	0.10	0.10	0.17	0.14	0.14	0.17	0.20
mut	1.00	0.96	0.98	0.95	0.95	0.95	0.74	0.85	0.85	0.80	0.71
nbC	13	11	12	10	10	10	6	7	7	6	5
emb	0.47	0.51	0.50	0.55	0.55	0.55	0.57	0.60	0.60	0.62	0.61
miTrue	0.92	0.90	0.93	0.89	0.89	0.89	0.70	0.79	0.79	0.74	0.65

Table 2.2: Comparative table ACF

the other methods find communities close to cliques (communities with higher density) and our method finds communities with more embeddedness (ratio of internal edges compared to the total number of edges).

In table 2.2, we can see that our algorithm was almost 100% accurate in reconstructing the true communities of the ACF (lost only to one instance of the Louvain method, by only 1%). Again, the label propagation method found communities very close to ours, but they are not the same, which proves that the methods are not equivalent. Now, on the other hand, our algorithm finds the communities with the lowest average size, highest average density and lowest average embeddedness. This suggests that, in the ACF example, our method finds communities closer to cliques compared to the other methods.

In both examples ZKC and ACF, our method uses the right criteria in order to find the network communities: our method discovers communities with higher density when the true communities have a high density value, and it discovers communities with higher embeddedness when the true communities have a high embeddedness value. For real-life networks it can also indicate that maybe the appropriate definition of communities is not the standard one finds in the literature, where communities are subsets with more edges inside them compared to edges outside them [255]; we could use the definition 3 as the standard definition of communities, where we only compare the edges inside a community with edges stemming from other communities, one at a time. Of course a much more detailed study must be conducted to shed light onto this phenomenon.

The reader should also notice that our definition of communities allows overlapping communities without any increase in complexity, which can be sometimes desirable (in figure 2.1 we found one node belonging to both communities).

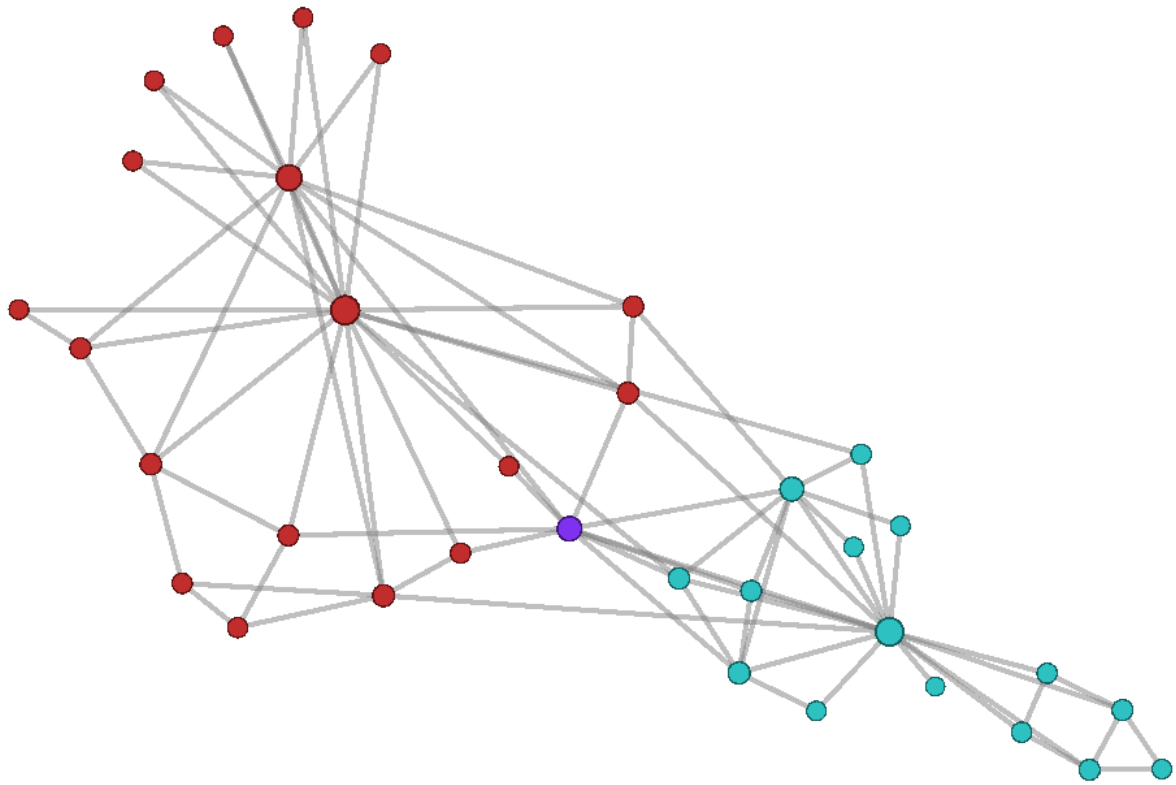

Figure 2.1: Communities for ZCK.

2.4.2 Dolphins

	OD	LP	L1	L2	L3	FM	FN	SRB1	SRB2	SRB3	R
mod	0.50	0.48	0.50	0.52	0.52	0.52	0.50	0.58	0.60	0.63	0.56
den	0.62	0.25	0.73	0.35	0.35	0.34	0.42	0.29	0.29	0.29	0.22
siz	0.14	0.33	0.10	0.20	0.20	0.20	0.25	0.25	0.25	0.25	0.33
mut	1.00	0.74	0.88	0.87	0.87	0.82	0.71	0.86	0.86	0.86	0.67
nbC	7	3	10	5	5	5	4	4	4	4	3
emb	0.47	0.75	0.40	0.56	0.56	0.56	0.58	0.65	0.65	0.65	0.72

Table 2.3: Comparative table Dolphins

We compare now our algorithm for the Dolphins network. Although we do not have its true communities, we can still make some remarks concerning the performance of our algorithm compared to the others, following table 2.3: first, one may see that contrary to the ZKC and ACF examples our

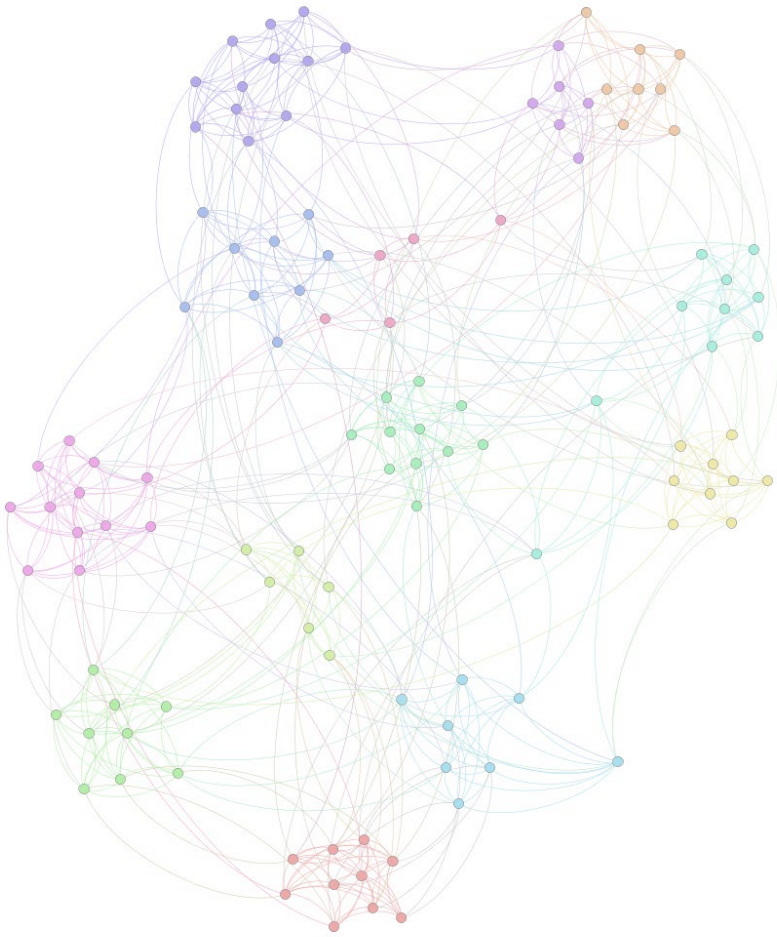

Figure 2.2: Communities for ACF.

algorithm found very different communities than the label propagation algorithm. Our algorithm has a number of communities between the first instance of the Louvain method and the other methods, keeping in the same order the modularity of the partitions and the average embeddedness.

However, we find communities with much higher average densities (together with the first instance of the Louvain method), which means that our communities resemble cliques more than the communities found by the other methods.

2.4.3 Facebook-ego

For the ten ego networks in Facebook, one can clearly see in table 2.4 that all methods found at least ten communities, with the exception of the greedy implementation using the Hamiltonian function of [260]. Of course the ego networks may have nested communities, which explains why there are almost systematically more than ten communities found by every method.

Again, the label propagation method gives very different results than our method, which resembles most the second instance of the Louvain method (the fact that our agorithm resembles one of the instances of the Louvain method seems to be fairly constant in other datasets).

Similarly to the Dolphins network example, our algorithm finds communities with the same

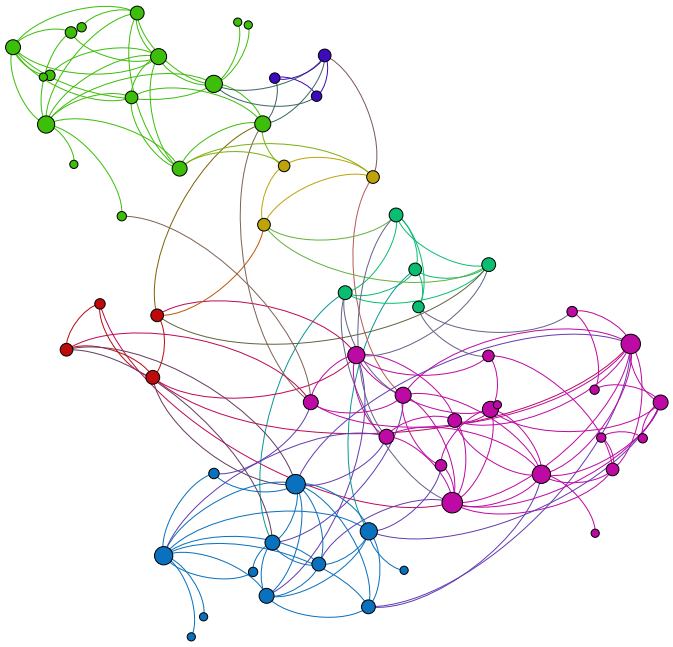

Figure 2.3: Communities for Dolphins.

	OD	LP	L1	L2	L3	FM	FN	SRB1	SRB2	SRB3	R
mod	0.79	0.16	0.81	0.83	0.83	0.83	0.78	0.86	0.87	0.89	0.77
den	0.33	0.04	0.68	0.31	0.31	0.29	0.37	0.18	0.18	0.18	0.23
siz	0.04	0.02	0.01	0.05	0.06	0.06	0.08	0.08	0.08	0.08	0.11
mut	1.00	0.39	0.80	0.81	0.78	0.80	0.72	0.80	0.80	0.80	0.70
nbC	25	51	101	19	17	16	13	13	13	13	9
emb	0.73	0.05	0.54	0.87	0.87	0.89	0.90	0.92	0.92	0.92	0.95

Table 2.4: Comparative table Facebook-ego

order of modularity and average embeddedness as the other methods, with higher than the average densities, which means that our communities resemble cliques more than the communities found by the other methods.

The lower average embeddedness may come from our definition of communities (definition 3), where we are concerned simply with pairwise comparison between communities. We do not attempt to minimize, for example, the number of edges between nodes inside and outside the communities,

as dictated by the standard definition of communities in [255].

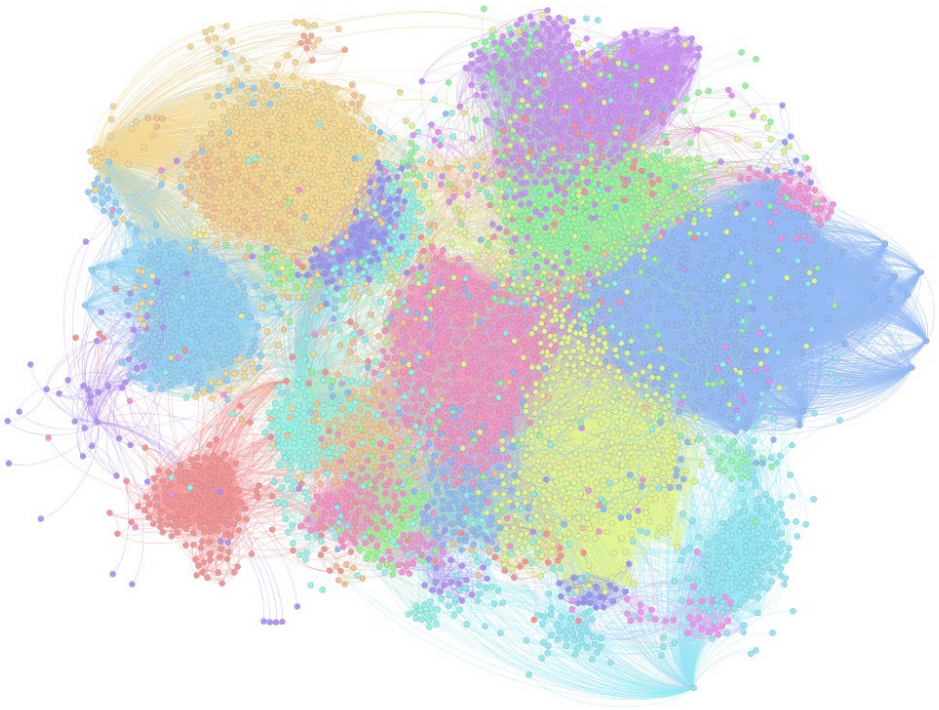

Figure 2.4: Communities for Facebook-ego.

2.4.4 GRQC-ArXiv

	OD	LP	L1	L2	L3	FM	FN	SRB1	SRB2	SRB3	R
mod	0.77	0.79	0.72	0.84	0.81	0.86	0.82	0.87	0.87	0.88	0.82
den	0.50	0.69	0.81	0.68	0.78	0.82	0.81	0.83	0.83	0.84	0.81
siz	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mut	1.00	0.93	0.92	0.87	0.77	0.76	0.74	0.75	0.75	0.73	0.71
nbC	633	726	1204	537	404	387	416	384	383	380	420
emb	0.69	0.80	0.58	0.90	0.94	0.98	0.97	0.99	0.99	0.99	0.97

Table 2.5: Comparative table GRQC-ArXiv

This example shows a result below the average when compared with other methods, where the communities discovered by our method have lower density (den) and embeddedness (emb) than the communities found by other methods. Our method finds communities resembling those of the label propagation and the Louvain method once again. However, one can see that the modularity of the communities found by our method still remains similar to other methods that are based on modularity-like optimization techniques.

This time, our method found a higher number of communities than the benchmark methods, which could explain the decrease in average density and embeddedness. This may be corrected by limiting the maximal number of communities to be found using the parameter K (number of contents in the underlying opinion dynamics algorithm)

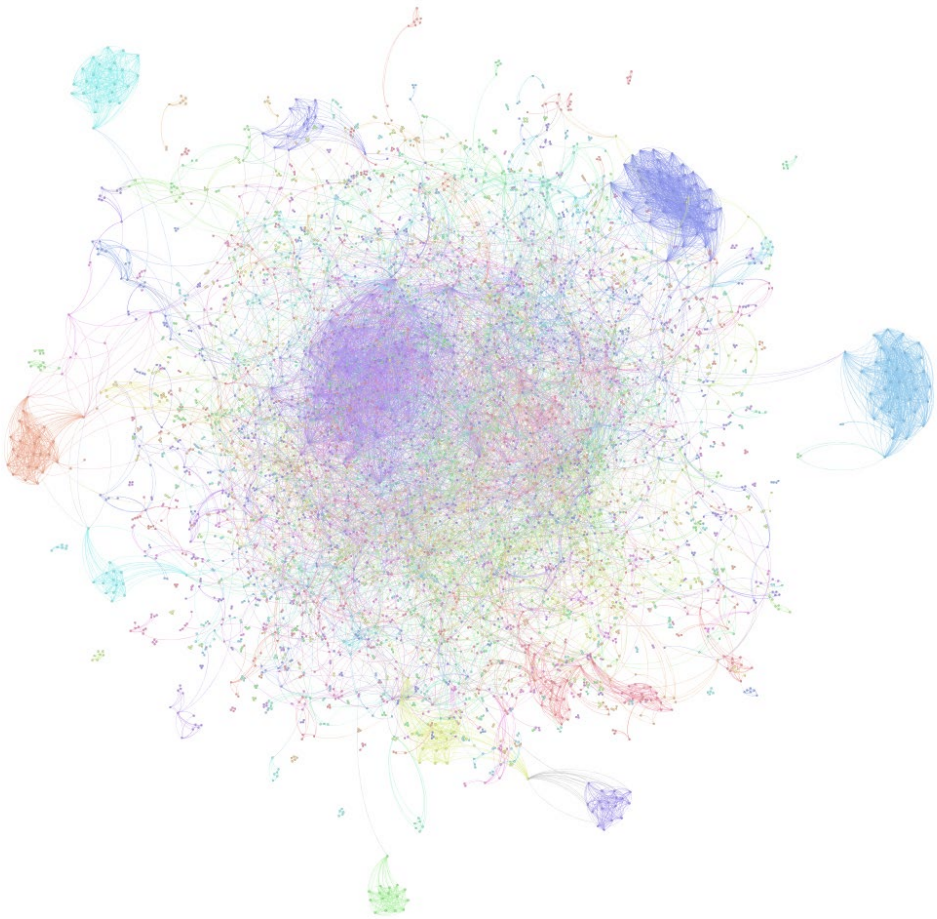

Figure 2.5: Communities for GRQC-ArXiv

2.4.5 Youtube

Similarly to the GRQC-ArXiv example, our method finds communities with lower density and embeddedness than other methods, with modularity still comparable to other methods that are based on modularity-like optimization techniques.

	OD	LP	L1	L2	L3	FM	SRB1	SRB2	SRB3
mod	0.52	0.38	0.64	0.68	0.61	0.67	0.69	0.70	0.72
den	0.17	0.50	0.66	0.37	0.47	0.50	0.59	0.59	0.63
siz	0.01	0.01	0.00	0.01	0.02	0.03	0.03	0.03	0.03
mut	1.00	0.42	0.64	0.57	0.47	0.49	0.45	0.45	0.42
nbC	141	73	628	79	52	40	33	33	30
emb	0.42	0.76	0.34	0.71	0.76	0.84	0.88	0.88	0.90

Table 2.6: Comparative table Youtube

Again, our method found a higher number of communities than the benchmark methods, which could explain the decrease in average density and embeddedness. Nevertheless, the unusual fact about this particular example is that *none* of the other methods seems to be similar to ours, which is given by the difference in normalized mutual information (mut) between the communities found by our method and the communities found by other methods.

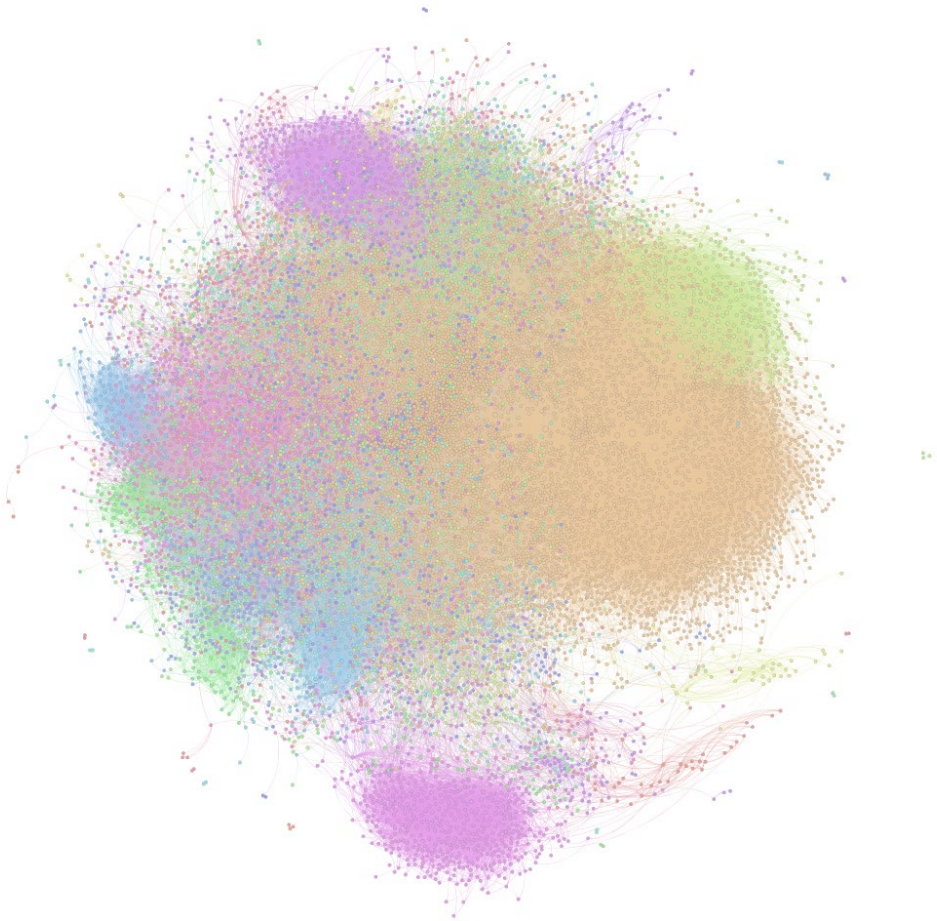

Figure 2.6: Communities for Youtube

2.5 Conclusion

We introduced in this chapter a new community detection algorithm based on a stochastic model of opinion dynamics. The proposed algorithm is studied using stochastic approximation techniques, which results in a precise description of the communities found.

In addition to the theoretical advantage over heuristic community detection methods, the presented algorithm is able to accommodate weighted networks, with the discovery of overlapping communities as an interesting byproduct with no mathematical or algorithmic overhead. Furthermore, one can add as well a priori information on the communities to be found, by choosing a suitable initial condition for the opinion dynamics algorithm.

Heuristic arguments for the fine-tuning of the parameters were presented, they establish two main implementations for the algorithm: a parametric one and a nonparametric one. The parametric version is less complex and caps the maximum number of communities to be found, whereas the non-parametric version does not make any assumptions on the maximum number of communities to be found. This choice gives the algorithm a multiscale character, allowing one to select the granularity of the unknown communities.

This algorithm has a manageable complexity and is also designed to be performed easily in a distributed fashion, making it useful for real-life networks.

Moreover, empirical tests with real-life benchmark graphs suggest that our less restrictive definition of communities could more easily fit real-life networks, where nodes inside communities do not need to be more linked within themselves, compared to the rest of the graph; it would suffice for them to be more linked only compared with other communities, taken one at a time.

Part II

Information Diffusion and Trend Detection

CHAPTER 3

Information diffusion using Hawkes processes

"If content is king, then conversion is queen."
— John Munsell, *CEO of Bizzuka*

3.1 Introduction

After studying in the first part of this thesis theoretical ideas of information diffusion on social networks by the means of opinion dynamics methods, we dedicate the second part of this thesis to a more computational and concrete approach to information diffusion. We derive in this chapter a general framework of information diffusion with Hawkes processes, and use one instance of this framework to create a trend detection algorithm.

Information diffusion/dissemination in social networks refers to users broadcasting (sharing, posting, tweeting, retweeting, liking, etc.) information to others in the network. By tweeting, for example, users broadcast information to the network, which is then transmitted to their followers. These sequences of broadcasts by users are called information cascades, and have been studied extensively over the past years; see for example [26, 174, 51]. The large amount of recent work on this subject reflects the strategic real-life implications which may be brought by the knowledge of such cascades: one can discover the hidden impact of users and contents on this diffusion, and highlight various characteristics of not only the social networks in question but also of the influential users and their contents [273, 126, 123].

Information cascades are complex objects, for which there is no consensus on the standard way to study them; for example: Kempe *et al.* in their seminal paper [169] develop a framework based on submodular functions to detect the optimal seed group in order to diffuse a fixed content in a social network, based on the so-called independent cascade propagation model [119, 120], which is a well known information diffusion model. In [231], Myers and Leskovec study the variation of the probability in retransmiting information due to previous exposure to different types of information; they found that, for Twitter, these retransmission probabilities are indeed very different when compared to results stemming from independent cascade models; however, their approach does not take into consideration the time between broadcasts of information and the topology of the network. And in [124], Gomez-Rodriguez *et al.* study the network inference problem from information cascades using survival theory; however, again, the authors to not take into consideration the underlying network structure.

Among the works dealing with information diffusion, there has been a steady increase of interest in point-processes-based models [310, 38, 158]. Point processes take into consideration the broadcast times of users, whereas a lot of information cascade models consider time to be discrete, i.e., time only evolves when events occur; point processes are counting processes and have thus a discrete state space, which makes them able to fully capture real-life features, such as the number of posts, without increasing the mathematical complexity of the models; and the closed formula for the likelihood of these point processes ([74] p. 232) gives us easy, simple and direct methods for the estimation of important model parameters. For instance, Myers *et al.* study in [232] the influence of externalities from other nodes on information cascades in networks; they use a point process approach, from which the time instances of infection are essential for the estimation of parameters, but the topological properties of the network are of secondary concern in their work.

One point process has been particularly useful in the modeling of these continuous-time models: the Hawkes process [140, 208]. Hawkes processes are self-exciting point processes and are perfect candidates for counting events on information cascades, where users transmit their information to their neighbors in a social network. The use of self-exciting processes here enlightens the necessity of a theory that can model the interaction between people having a conversation or exchanging messages: imagine two people messaging each other through SMS. Normally each one would have its own rhythm of messaging, but due to the self-excitation among these people, they will text and respond faster than they would normally do when generating SMS messages without response. For example, Yang and Zha study in [310] the propagation of memes (see definition in [79] p. 192.) in social networks with linear Hawkes processes and couple the point process with a language model in order to estimate the memes. They provide a variational Bayes algorithm for the coupled estimation of the language model, the influence of users and their intrinsic diffusion rates; however, they do not take into consideration the influence that memes may have on one another; moreover, they propose the estimation of the entire social network, not taking into consideration the eventual lack of communication between users.

Hawkes processes have already been successfully used to study earthquakes [242], neuronal activities [40], high-frequency finance [15], social sciences [201, 71] and many other fields, with a vast and diversified literature.

This chapter aims to provide a solid and rich framework for information diffusion models in social networks using Hawkes processes. The presented framework is capable of:

- modeling and estimating *user-user* and *topic-topic* interactions,
- modeling and estimating *multiple* social networks and their interactions,
- being combined with topic models [34, 265, 33], for which modified collapsed Gibbs sampling [77, 128] and variational Bayes techniques [151, 145] are derived,
- estimating *different temporal effects* of the users diffusion, such as seasonality and non-homogeneity,
- using and estimating *dynamic/temporal* social networks [154], and
- retrieving the *community structure* of the underlying users influence in social networks, due to a dimensionality reduction during the parameters estimation (see [182] for another example of such methodology).

This chapter is organized as follows. Section 3.2 describes the models for our Hawkes information diffusion framework. Section 3.3 details the estimation procedure of the hidden influences. Section

3.4 discusses some additional topics for the Hawkes diffusion framework. In Section 3.5 numerical experiments are performed with four different datasets, and Section 3.6 concludes the chapter.

3.2 Hawkes diffusion models

We start with brief introduction to Hawkes processes: a multivariate linear Hawkes process (see [140, 208] for more details) is a self-exciting orderly point process $X_t,\ t \in [0, \tau]$ with intensity $\lambda_t = \lim_{\delta \searrow 0} \frac{\mathbb{E}[X_{t+\delta} - X_t | \mathcal{F}_t]}{\delta}$ satisfying

$$\lambda_t = \mu + \int_0^t \phi(t-s) dX_s,$$

where $\mathcal{F}_t = \sigma(X_s,\ s \leq t)$ is the filtration generated by X, μ is an intrinsic Poissonian rate and ϕ is a causal kernel that is responsible for the self-exciting part.

The Hawkes intensity λ_t can be divided into two distinguished parts: the intrinsic Poissonian rate μ, which models the base intensity of the Hawkes process, and does not take into account the past of the process, and the self-exciting part $\int_0^t \phi(t-s) dX_s$, which models the interactions of the present with past events. The μ coefficient can, for example, model how some user tweets something, after learning about it in class or at work, after listening to the radio or watching television.

The orderly property of the Hawkes process means that X cannot have two events/jumps at the same time ([74] p. 232), and by the standard theory of point processes ([74] p. 233) we have that an orderly point process is completely characterized by its intensity, which in this case is also a stochastic process.

The self-excitatory property of the Hawkes process means that for all $0 \leq u < t < s \leq \tau$ we have that: for every pair of coordinates (i, j)

$$cov(X_s^i - X_t^i, X_t^j - X_u^j) \geq 0,$$

and there exists at least one pair of coordinates (i^*, j^*) such that

$$cov(X_s^{i^*} - X_t^{i^*}, X_t^{j^*} - X_u^{j^*}) > 0,$$

This means that the future jumps of a self-exciting process become more probable to occur when the process jumps.

We place ourselves under a generic framework for information diffusion: N users of a social network disseminate their information over a social network. The dissemination/broadcasting of messages can be performed in various ways, depending on the application and the social network in question: measuring tweets or retweets, checking the history of a conversation in a chat room, "pinning" pictures, etc. However, they all have one thing in common: messages are broadcasted by the N users in the social network, and those users that can receive these broadcasts are influenced by them (at least indirectly).

The social network is defined as a communication graph $G = (V, E)$, where V is the set of users and E is the edge set, i.e., the set with all the possible communication links between users of the social network. We assume the graph to be directed and unweighted, and coded by an inward adjacency matrix A such that $A_{i,j} = 1$ if user j is able to broadcast messages to user i, or $A_{i,j} = 0$ otherwise. We define $j \rightsquigarrow i$ if and only if $A_{i,j} = 1$, i.e., if i can be influenced by user j.

Users influence each other when broadcasting, in the following way: when user i broadcasts something, all users that receive the broadcasts from user i (users j such that $A_{j,i} = 1$) see this

broadcast. They are then more compelled to reiterate this procedure in the future, by broadcasting on their own. Let us take Twitter for example: user i tweets something in Twitter, and thus users that follow user i receive this tweet on their walls. They read this tweet and may become inclined to answer it, comment it, rebuke it, or even retweet it. This means that this tweet from user i created a cascade effect, which provided all his followers an increase in their tweeting probability on the future. This signifies that user i influenced his followers.

Throughout this chapter we adopt two kinds of message categories: the first kind assumes that messages are of K predefined topics (economics, religion, culture, politics, sports, music, etc.) and that each message is represented by exactly one of these topics. The second kind assumes a "fuzzy" setup with K topics, where this time the topics are not known beforehand and messages are a mixture of these unknown (latent) topics; Barack Obama may for example tweet something that has 40% of its content about politics, 50% of its content about economics and 10% of its content related to something else.

3.2.1 User-user and topic-topic interactions with predefined topics

We first focus on the case where messages are about one of K predefined contents (economics, religion, sports, etc.). We assume that we have N users in a single social network and that they can influence each other to broadcast, and that these influences are independent of the broadcasted content. On the other hand, the topic to be broadcasted is influenced by the topics already broadcasted beforehand.

This is the case, for example, if one wants to separate the influence effects of users and topics: posts about politics can influence posts about fashion, economics, religion, etc., and people can influence other people simply because they are friends, famous or charismatic. In this model we assume that the influence of a specific user when posting something is given by two different components, the user-user component and the topic-topic component.

These influences are coded by two matrices J and B such that $J_{i,j} \geq 0$ is the influence of user i over user j and $B_{c,k} \geq 0$ is the influence of topic c over topic k.

We model the number of messages broadcasted by users as a linear Hawkes process $X_t \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$, where $X_t^{i,k}$ is the cumulative number of messages of topic k broadcasted by user i until time $t \in [0,\tau]$ in the social network. In other words, this X_t is a $\mathbb{R}^{N\times K}$ point process with intensity

$$\lambda_t^{i,k} = \mu^{i,k} + \sum_c \sum_{j \rightsquigarrow i} B_{c,k} J_{i,j} \int_0^{t-} \phi(t-s) dX_s^{j,c} = \mu^{i,k} + \sum_c \sum_{j \rightsquigarrow i} B_{c,k} J_{i,j} (\phi * dX)_t^{j,c},$$

where $\mu^{i,k} \geq 0$ is the intrinsic rate of broadcasting of user i about topic k, $\phi(t) \geq 0$ is the temporal influence kernel that measures the temporal shape of influences coming from past broadcasts - which satisfies $||\phi||_1 = \int_0^\infty \phi(t)dt < \infty$ - and

$$(\phi * dX)_t = \int_0^{t-} \phi(t-s) dX_s \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$$

is the convolution matrix of the temporal kernel ϕ and the jumps dX. This allows one to use $N^2 + K^2$ parameters instead of N^2K^2 for the full fledged model without this influence factorization.

As said before, not all users can communicate among themselves. Hence one must take into consideration the inward adjacency matrix A given by the underlying structure on the social network. This is done by the relation

$$A_{i,j} = 0 \Rightarrow J_{i,j} = 0. \tag{3.1}$$

Remark: Two standard time-decaying functions are $\phi(t) = \omega e^{-\omega t}.\mathbb{I}_{\{t>0\}}$ a light-tailed exponential kernel [245, 246, 310] and $\phi(t) = (b-1)(a+t)^{-b}.\mathbb{I}_{\{t>0\}}$ a heavy-tailed power-law kernel (see [71]). Expectation-minimization algorithms can be derived in order to estimate the parameters ω in the exponential case [133, 198] and a, b in the power-law case, as provided in appendix B.

3.2.2 User-topic interactions and global influence in the social network

A different model arises when users do not influence other individually, but they influence the social network as a whole. This means that instead of having an influence matrix $J \in \mathcal{M}_{N\times N}(\mathbb{R}_+)$ that measures the user-user interactions, we have now an influence matrix $\tilde{J} \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$ such that $\tilde{J}_{i,k} \geq 0$ is the influence of user i over the whole social network, when he broadcasts something about topic k.

Hence, the associated Hawkes process $X_t^{i,k}$, which measures the cumulative number of messages broadcasted by user i about topic k until time $t \in [0,\tau]$, has intensity

$$\begin{aligned}\lambda_t^{i,k} &= \mu^{i,k} + \sum_c \sum_{j \rightsquigarrow i} B_{c,k}\tilde{J}_{j,c} \int_0^{t-} \phi(t-s)dX_s^{j,c} \\ &= \mu^{i,k} + \sum_c \sum_j B_{c,k}A_{i,j}\tilde{J}_{j,c} \int_0^{t-} \phi(t-s)dX_s^{j,c},\end{aligned}$$

Think about Barack Obama: it is natural that posts or tweets about economics or politics coming from Obama are going to have a much bigger impact than posts about sports or fashion. People normally have the most influence in their areas of expertise, and we develop a model that accommodates this feature.

3.2.3 User-user and topic-topic interactions with "fuzzy" topic label

Up until now we have dealt with information dissemination models having K predefined topics and in which each broadcasted message was assumed to belong to one, and only one, of these topics. We consider now a different point of view regarding the broadcasted messages: each message now is a mixture over K undiscovered/latent topics. These topics are distributions over words and each message broadcasted at time $t_s \in [0,\tau]$ generates the *message's empirical distribution of topics* random variable Z^{t_s} such that

$$Z_k^{t_s} = \frac{1}{N_s}\sum_{w=1}^{N_s} z_k^{s,w}, \tag{3.2}$$

where N_s is the number of words in the message broadcasted at time t_s and $z^{s,w}$ are independent discrete random variables modeling the topic of word w, i.e., $z_k^{s,w} = 1$ if and only if word w at message t_s is about topic k, and 0 otherwise.

In this model users receive messages that are mixtures of topics and each user reacts to topics in a different manner, these user-topic interactions are characterized by the matrix $b \in \mathcal{M}_{N\times K}(\mathbb{R}_+)$, such that $b_{i,k}$ measures the influence of topic k over user i.

We define thus the Hawkes processes X_t^i as the cumulative number of messages broadcasted by

user i in the social network until time $t \in [0, \tau]$, with intensity

$$\lambda_t^i = \mu^i + \sum_{j \rightsquigarrow i} J_{i,j} \sum_{c,k} B_{c,k} b_{i,k} \int_0^{t-} \phi(t-s) Z_c^s dX_s^j$$
$$= \mu^i + \sum_{j \rightsquigarrow i} J_{i,j} \sum_{c,k} B_{c,k} b_{i,k} (\phi *_Z dX)_t^{j,c},$$

where $\mu^i \geq 0$ represents the intrinsic dissemination rate of user i and

$$(\phi *_Z dX)_t^{j,c} = \int_0^{t-} \phi(t-s) Z_c^s dX_s^j$$

is the (j, c) entry of the weighted convolution of the temporal kernel ϕ and the jumps dX, where the weights are the topic empirical proportions of each message broadcasted by user j.

Again, not all users can communicate among themselves, hence one must take into consideration Eqn. (3.1).

In order to fully exploit the random variables Z^{t_s} we use topic models [34, 265, 33], as for example the latent Dirichlet allocation [34] (see for [211] such a methodology) or the author-topic model [265]. More details about topic models can be found in appendix C.

Remark: One can also easily extend the model in subsection 3.2.2 to the "fuzzy" diffusion framework, following these ideas.

3.2.4 User-user and topic-topic interactions with predefined topics in multiple social networks

We now turn to the case where we have M "interconnected" social networks. The m^{th} social network is defined as a communication graph $G^m = (V^m, E^m)$, where V^m is the set of users and E^m is the edge set, i.e., the set with all the possible communication links between users of the m^{th} social network. We assume these graphs to be directed and unweighted, and coded by inward adjacency matrices A^m such that $A_{i,j}^m = 1$ if user j is able to broadcast messages to user i on social network m, or $A_{i,j}^m = 0$ otherwise. Having this collection of unweighted inward adjacency matrices $(A^m)_{m \in \{1,2,\cdots,M\}}$, we define $j \overset{m}{\rightsquigarrow} i$ if and only if $A_{i,j}^m = 1$, i.e., if i can be influenced by user j through social network m.

One can think about Facebook and Twitter users: there are users in Facebook that do not necessarily follow the same people on Facebook and on Twitter, and vice-versa. The explanation is quite simple: Facebook posts are of a different nature than Twitter posts, thus the following process on both networks is also different. Let us say that Facebook is social network 1 and Twitter is social network 2; $A_{i,j}^1 = 1$ means that user i follows user j in Facebook and receives the news published by user j in his or her timeline. As said, that does not necessarily imply that $A_{i,j}^2 = 1$, i.e., user i also follows user j on Twitter.

This network formalism can be associated with the multiplex network formalism [87, 172]: *multiplex networks* (or "multirelational" networks) are networks where links have different characteristics, thus one node may have more than one edge linking it to another node. It may also happen that nodes do not have all types of links, only a few of them (maybe even none). In our case, specifically, we may consider the ensemble of M social networks as a multiplex network with node set $V = \bigcup_m V^m$ (with cardinality $N = \sharp(\bigcup_m V^m)$) and edge set $E = \bigcup_m E^m$, and links being characterized by the social network in question, i.e., each social network m has its own set of links E^m,

and two users (nodes) may have multiple links between them, each of them associated to a different social network.

Users influence each other when broadcasting, in the following way: when user i broadcasts something on social network m, all users that receive the broadcasts from user i (users j such that $A^m_{j,i} = 1$) see this broadcast. They are then more compelled to reiterate this procedure in the future, by broadcasting on their own, which may not necessarily happen on the same social network.

Assuming that we have M different social networks, each one with its own adjacency matrix A^m, we model the influence of broadcasts using, similarly to the model in subsection 3.2.1, three matrices $J \in \mathcal{M}_{N\times N}(\mathbb{R}_+)$, $B \in \mathcal{M}_{K\times K}(\mathbb{R}_+)$ and $S \in \mathcal{M}_{M\times M}(\mathbb{R}_+)$, such that $J_{i,j} \geq 0$ is the influence of user i over user j, $B_{c,k} \geq 0$ is the influence of topic c over topic k and $S_{m,n}$ is the influence that a generic user of the social network m has over a generic user of the social network n. The network-network influence matrix S measures thus how broadcasts made on one social network influence broadcasts made on the others.

Let $X^{i,k,n}_t$ be the cumulative number of messages broadcasted by user i about content k at social network n until time $t \in [0, \tau]$. The intensity for this process is thus

$$\lambda^{i,k,n}_t = \mu^{i,k,n} + \sum_{m,c,j\rightsquigarrow i} S_{m,n} J_{i,j} B_{c,k} \int_0^t \phi^n(t-s) dX^{j,c,m}_s,$$

where J is again the user-user influence matrix, B is the topic-topic influence matrix and μ is the intrinsic rate of dissemination on different social networks.

In view of Eqn. (3.1), if there exists an edge $j \rightsquigarrow i$ in some social network, then user i can be influenced by user j. Our new constraint becomes

$$\sum_m A^m_{i,j} = 0 \Rightarrow J_{i,j} = 0.$$

One can notice in the definition for the intensity of this model that each social network m has its own[1] temporal kernel function ϕ^m. Each temporal kernel ϕ^m represents how users and contents in each social network are affected by ancient messages, and are considered a *timescale* parameter[2]. Let us take for comparison Twitter and Flickr: in Twitter users chat, discuss, posts comments and retweets, while Flickr is a photo-sharing social network that allows users to upload photos and post comments. This means that the conversation and interaction mechanisms in both social networks are different, since they serve different purposes. It is thus natural to assume that users in both social networks react differently to the information received; these different reactions are in part measured by the different temporal kernels $(\phi^m)_{m\in\{1,\cdots,M\}}$.

Remark: One can notice that this factorization of influences allows us to use $N^2 + K^2 + M^2$ parameters instead of $N^2K^2M^2$, which decreases in a great amount the complexity of the system and the estimation time.

3.2.5 Network dependent user-user and topic-topic interactions in multiple social networks

A second (and more complex) extension to the single social network information diffusion model is to assume that the different broadcasting mechanisms in each social network imply different

1. The temporal kernel functions could take more complicated forms, such as $\phi^{k,m}$, where each topic in a social network would have an idiosyncratic temporal kernel function. This enlightens the versatility of this Hawkes framework, allowing one to adapt the system parameters to any desired situation.
2. Take for example the exponential kernel $\phi(t) = \omega e^{-\omega t}.\mathbb{I}_{\{t>0\}}$: the larger the ω, the larger is the influence of recent broadcasts. This may imply users responding faster to immediate messages.

influences on users and topics. It means that the user-user and topic-topic influences are now specific to each social network, i.e., user j broadcasting a message about content c on a social network m influences user i *in this same social network* when he broadcasts some message about content k. These network-dependent influences are measured by the user-user influence matrices $(J^m)_{m\in\{1,\cdots,M\}}$ and topic-topic influence matrices $(B^m)_{m\in\{1,\cdots,M\}}$.

Remark: Viewed as high-dimensional objects, J and B are three-dimensional tensors.

We can define, again, $X_t^{i,k,n}$ to be the cumulative number of messages broadcasted by user i about content k at social network n until time $t \in [0,\tau]$. The intensity for this process is then

$$\lambda_t^{i,k,n} = \mu^{i,k,n} + \sum_{m,c,j \overset{m}{\rightsquigarrow} i} S_{m,n} J_{i,j}^m B_{c,k}^m \int_0^t \phi^n(t-s) dX_s^{j,c,m},$$

where $j \overset{m}{\rightsquigarrow} i$ means that user j can influence user i in social network m, i.e., $A_{i,j}^m = 1$.

Since now users only influence themselves in the same social network, the adjacency matrix constraint in Eqn. (3.1) becomes

$$A_{i,j}^m = 0 \Rightarrow J_{i,j}^m = 0.$$

Remark: One can easily extend the model with social network-social network specific influences of the form $J_{i,j}^{m,n}$ and $B_{c,k}^{m,n}$, for which the above extension is a particular case $J_{i,j}^{m,n} = J_{i,j}^m S_{m,n}$ and $B_{c,k}^{m,n} = B_{c,k}^m S_{m,n}$.

Remark: One can also easily extend the model in subsections 3.2.2 and 3.2.3 to take into account multiple social networks, following the same ideas.

3.2.6 General interaction model with predefined topics in multiple social networks

We provide, for the sake of completeness, the most general model of interactions in social networks. It occurs when one does not factorize the interactions of users, contents and networks, as in the previous cases. The influences have now the full form $\Gamma_{(i,k,n)}^{(j,c,m)}$, where $\Gamma_{(i,k,n)}^{(j,c,m)}$ measures the influence of the user-content-network triple (j,c,m) on the user-content-network triple (i,k,n), i.e., how a broadcast of user j about content c in social network m influences a broadcast of user i about content k in social network n.

The intensity of a model with predefined topics takes the form

$$\lambda_t^{i,k,n} = \mu^{i,k,n} + \sum_m \sum_c \sum_{j \overset{m}{\rightsquigarrow} i} \Gamma_{(j,c,m)}^{(i,k,n)} \int_0^{t-} \phi^n(t-s) dX_s^{j,c,m}$$

and have all other models as particular cases.

The usefulness of the simpler models regards the number of hidden influence parameters to be estimated: for the full general model one has $N^2K^2M^2$ parameters to estimate, whereas in the simpler ones one only has, for example, at most $N^2 + K^2 + M^2$ in subsection 3.2.4 and at most $M(N^2 + K^2)$ in subsection 3.2.5.

3.3 Maximum likelihood estimation and multiplicative updates

Section 3.2 describes different parametric models of information diffusion using Hawkes processes, all of them exploiting different peculiarities of the reality. One of the strong points about

point processes (and Hawkes processes for that matter) is the analytic form of the likelihood of their realization (see [240] or [74] p. 232), where Hawkes-based models for information diffusion used extensively this property in order to derive convex-optimization-based maximum likelihood estimates for the system parameters [158, 310, 317]. For example, in our Hawkes diffusion framework, we may estimate the user-user influence matrix J, the content-content influence matrix B, the network-network influence matrix S, the users intrinsic dissemination rate μ, etc.

A different technique for the maximum likelihood estimation of the Hawkes process X was derived in [245, 246], where the authors slice the information time period $[0, \tau]$ into T small bins of size $\delta > 0$ in order to create suitable tensors for the intensity λ_t and the Hawkes jumps dX_t, and show that maximizing an approximation of the log-likelihood is equivalent to solving a nonnegative[3] tensor factorization (NTF) problem [192, 170, 66]. This section is thus dedicated to demonstrating that all information dissemination models in section 3.2 can be estimated using the same techniques, which creates an unified information dissemination framework using Hawkes processes and topic models.

Since we deal with real-life social networks, the number of parameters to be estimated is large and convex optimization techniques that estimate each parameter separately are too demanding in terms of complexity. That is why we adopt the estimation framework of [245, 246], for which multiplicative updates[4] can be derived (see [212, 211] for the same methodology).

Let us take a $\delta > 0$ that is smaller than the minimum elapsed time between broadcasts in $[0, \tau]$ and divide $[0, \tau]$ into $T = \lceil \frac{\tau}{\delta} \rceil$ time bins such that we do not have more[5] than one broadcast in each bin, in order to preserve the orderliness property of X.

Let Y, $\overline{\lambda}$ and $\overline{\phi}$ be tensors such that

$$Y_t = \frac{dX_{(t-1)\delta}}{\delta} = \frac{X_{t\delta} - X_{(t-1)\delta}}{\delta}, \qquad \overline{\lambda}_t = \lambda_{(t-1)\delta} \quad \text{and}$$

$$\overline{\phi}_t^m = \begin{cases} (\phi^m * dX)_{(t-1)\delta} & \text{for predefined topics model} \\ (\phi^m *_Z dX)_{(t-1)\delta} & \text{for "fuzzy" diffusion model,} \end{cases}$$

i.e. Y contains the jumps of X_t at each time bin $((t-1)\delta, t\delta]$.

We begin our estimation procedure by showing that maximizing the Riemann-sum approximation of the log-likelihood of X is equivalent to minimizing the Kullback-Leibler (KL) divergence between Y and $\overline{\lambda}$.

Lemma 13. *If $\int_0^\tau \log(\lambda_t^{i,k,m}) dX_t^{i,k,m}$ and $\int_0^\tau \lambda_t^{i,k,m} dt$ are approximated by their respective Riemann sums, then maximizing the approximated log-likelihood of X in $[0, \tau]$ is equivalent to minimizing*

$$D_{KL}(Y|\overline{\lambda}) = \sum_{i,k,m,t} d_{KL}(Y_t^{i,k,m} | \overline{\lambda}_t^{i,k,m}), \tag{3.3}$$

where $d_{KL}(y|x) = y \log(\frac{y}{x}) - y + x$ is the Kullback-Leibler divergence between x and y.

Proof. Let us place ourselves, without loss of generality, in an information diffusion model with

3. By nonnegative we mean tensors with nonnegative entries.
4. The multiplicative updates using NTF techniques are only one of the existing estimation techniques. Alternative methods are discussed in subsection 3.4.3.
5. In practice, this orderliness constraint is not satisfied in order to decrease the complexity of the multiplicative updates.

predefined topics[6] and let t_n be the broadcast instants in $[0,\tau]$, such that user i_n broadcasted a message about topic k_n in social network m_n at time t_n, i.e., t_n is the n^{th} broadcasting time in the M social networks.

We have that the log-likelihood of X is given by (see for example [240] or [74] p. 232)

$$\begin{aligned}\mathcal{L} &= \log\left(\prod_{0\leq t_n\leq\tau}\lambda_{t_n}^{i_n,k_n,m_n}\right) - \sum_{i,k,m}\int_0^\tau \lambda_t^{i,k,m}dt \\ &= \sum_{i,k,m}\left(\int_0^\tau \log\lambda_t^{i,k,m}dX_t^{i,k,m} - \int_0^\tau \lambda_t^{i,k,m}dt\right).\end{aligned}$$

Approximating the integrals in $\mathcal{L}$ by their Riemann sums we get

$$\mathcal{L} \sim \sum_{i,k,m}\sum_t\left(\log\lambda_{(t-1)\delta}^{i,k,m}(X_{t\delta}^{i,k,m} - X_{(t-1)\delta}^{i,k,m}) - \delta\lambda_{(t-1)\delta}^{i,k,m}\right),$$

thus maximizing the approximation of $\mathcal{L}$ is equivalent to minimizing

$$-\mathcal{L}/\delta \sim \sum_{i,k,m}\sum_t\left(\overline{\lambda}_t^{i,k,m} - Y_t^{i,k,m}\log\overline{\lambda}_t^{i,k,m}\right).$$

With Y fixed, this is equivalent to minimizing

$$D_{KL}(Y|\overline{\lambda}) = \sum_{i,k,m,t} d_{KL}(Y_t^{i,k,m}|\overline{\lambda}_t^{i,k,m}).$$

□

Using lemma 13, we have that the maximization of the approximated log-likelihood of X is equivalent to a nonnegative tensor factorization problem with cost function $D_{KL}(Y|\overline{\lambda})$, where Y are the normalized jumps of X and $\overline{\lambda}$ is a tensor representing the intensity of X.

This nonnegative tensor factorization problem stemming from the minimization of the cost function $D_{KL}(Y|\overline{\lambda})$ has already been studied at length in [192, 193, 170], where authors derive convergent multiplicative updates [176, 100].

These multiplicative updates are interesting for several reasons: they are simple to implement (they are basically matrix products and entrywise operations), can be performed in a distributed fashion and have a low complexity on the data, thus being adequate to work on real-life social network of millions (or even hundreds of millions) of nodes.

These NTF techniques are based on the multiplicative updates given by the following lemma:

Lemma 14. *Let Y be a nonnegative tensor of dimension M, S a nonnegative tensor of dimension $s_S + L$ and H a nonnegative tensor of dimension $h_H + L$ such that $s_S + h_H \geq M$. Define SH, a the nonnegative tensor of dimension M, such that*

$$(SH)_{j_1,\cdots,j_M} = \sum_{l_1,\cdots,l_L} S_{i_{s_1},\cdots,i_{s_S},l_1,\cdots,l_L} H_{i_{h_1},\cdots,i_{h_H},l_1,\cdots,l_L},$$

where we have that

6. For "fuzzy" diffusion models, we consider the conditional log-likelihood with respect to Z, which is (see for example [74] p. 251)

$$\mathcal{L}(X|Z) = \log\left(\prod_{0\leq t_n\leq\tau}\lambda_{t_n}^{i_n,m_n}\right) - \sum_{i,m}\int_0^\tau\lambda_t^{i,m}dt = \sum_{i,m}\left(\int_0^\tau\log\lambda_t^{i,m}dX_t^{i,m} - \int_0^\tau\lambda_t^{i,m}dt\right).$$

- $\{i_{s_1}, \cdots, i_{s_S}\} \cup \{i_{h_1}, \cdots, i_{h_H}\} = \{j_1, j_2, \cdots, j_M\}$ *(we can still have* $\{i_{s_1}, \cdots, i_{s_S}\} \cap \{i_{h_1}, \cdots, i_{h_H}\} \neq \emptyset$*) and*
- $\{j_1, \cdots, j_M\} \cap \{l_1, \cdots, l_L\} = \emptyset$.

Define the cost function

$$D_{KL}(Y|SH) = \sum_{j_1,\cdots,j_M} d_{KL}(Y_{j_1,\cdots,j_M}|(SH)_{j_1,\cdots,j_M}),$$

where $d_{KL}(y|x) = y\log(\frac{y}{x}) - y + x$ *is the Kullback-Leibler divergence between* x *and* y.

The multiplicative updates for $D_{KL}(Y|SH)$ *of the form*

$$Z^{n+1} \leftarrow Z^n \odot \frac{\nabla_Z^- D_{KL}(Y|SH)_{|Z^n}}{\nabla_Z^+ D_{KL}(Y|SH)_{|Z^n}}, \tag{3.4}$$

with

- *the variables* $Z \in \{S, H\}$, $\nabla_Z^{+/-} D_{KL}(Y|SH)$ *the positive/negative part of* $\nabla_Z D_{KL}(Y|SH)$,
- $A \odot B$ *the entrywise product between two tensors* A *and* B, *and*
- $\frac{A}{B}$ *the entrywise division between two tensors* A *and* B,

satisfy

$$D_{KL}(Y|S^{n+1}H) \leq D_{KL}(Y|S^nH) \quad \text{and} \quad D_{KL}(Y|SH^{n+1}) \leq D_{KL}(Y|SH^n),$$

i.e., the multiplicative updates produce nonincreasing values for the cost function $D_{KL}(Y|SH)$.

Proof. We prove the result only for the tensor S, the calculations for the tensor H are equivalent. Let

$$D_{KL}(Y|SH) = \sum_{j_1,\cdots,j_M} d_{KL}(Y_{j_1,\cdots,j_M}|(SH)_{j_1,\cdots,j_M}),$$

where $d_{KL}(y|x) = y\log(\frac{y}{x}) - y + x$ is the Kullback-Leibler divergence between x and y.

In order to find suitable multiplicative updates for this cost function we proceed in the same manner as in [100, 176], i.e., we find an auxiliary function G such that $G(S, \tilde{S}) \geq D(Y|SH)$ for all nonnegative tensor S and $G(S, S) = D(Y|SH)$, with the NTF updates S^n, $n \in \{0, 1, 2, \cdots\}$ of the form

$$S^{n+1} = argmin_{X \geq 0} G(X, S^n). \tag{3.5}$$

We have thus

$$\begin{aligned} D(Y|S^{n+1}H) &\leq G(S^{n+1}, S^n) = \min_{\tilde{S} \geq 0} G(\tilde{S}, S^n) \\ &\leq G(S^n, S^n) = D(Y|S^nH). \end{aligned}$$

Let $\mathcal{J} = \{j_1, \cdots, j_M\}$, $\mathcal{S} = \{i_{s_1}, \cdots, i_{s_S}\}$, $\mathcal{H} = \{i_{h_1}, \cdots, i_{h_H}\}$ and $\mathcal{L} = \{l_1, \cdots, l_L\}$ be the index sets for the tensor summations such that $\mathcal{S} \cup \mathcal{H} = \mathcal{J}$ and $\mathcal{J} \cap \mathcal{L} = \emptyset$, and define the function G as

$$G(S, \tilde{S}) = \sum_{\mathcal{J}} \sum_{\mathcal{L}} \frac{S_{\mathcal{S},\mathcal{L}} H_{\mathcal{H},\mathcal{L}}}{\tilde{Y}_{\mathcal{J}}} d_{KL}(Y_{\mathcal{J}}|\tilde{Y}_{\mathcal{J}} \frac{S_{\mathcal{S},\mathcal{L}}}{\tilde{S}_{\mathcal{S},\mathcal{L}}})$$

where $\tilde{Y}_{\mathcal{J}} = \sum_{\mathcal{L}} \tilde{S}_{\mathcal{S},\mathcal{L}} H_{\mathcal{H},\mathcal{L}}$ (if $\tilde{S} = S$ then $\tilde{Y}_{\mathcal{J}} = \sum_{\mathcal{L}} S_{\mathcal{S},\mathcal{L}} H_{\mathcal{H},\mathcal{L}} = (SH)_{\mathcal{J}}$).

We easily have that $G(S,S) = D(Y|SH)$. Moreover, by the convexity of $d_{KL}(x|y)$ in y and $\sum_{\mathcal{L}} \frac{\tilde{S}_{\mathcal{S},\mathcal{L}} H_{\mathcal{H},\mathcal{L}}}{\tilde{Y}_{\mathcal{J}}} = 1$, we have that

$$\begin{aligned} G(S,\tilde{S}) &\geq \sum_{\mathcal{J}} d_{KL}(Y_{\mathcal{J}} | \sum_{\mathcal{L}} \frac{\tilde{S}_{\mathcal{S},\mathcal{L}} H_{\mathcal{H},\mathcal{L}}}{\tilde{Y}_{\mathcal{J}}} \tilde{Y}_{\mathcal{J}} \frac{S_{\mathcal{S},\mathcal{L}}}{\tilde{S}_{\mathcal{S},\mathcal{L}}}) \\ &= \sum_{\mathcal{J}} d_{KL}(Y_{\mathcal{J}} | \sum_{\mathcal{L}} S_{\mathcal{S},\mathcal{L}} H_{\mathcal{H},\mathcal{L}}) = D(Y|SH), \end{aligned}$$

thus G is indeed an auxiliary function.

Now, we calculate the multiplicative updates for this auxiliary function as in Eqn. (3.5). Taking the gradient $\nabla_S G(S^{n+1}, S^n) = 0$ gives us

$$\begin{aligned} \partial_{S_{\mathcal{S},\mathcal{L}}} G(S^{n+1}, S^n) &= \sum_{\mathcal{J} \setminus \mathcal{S}} \left(1 - \frac{Y_{\mathcal{J}} S^n_{\mathcal{S},\mathcal{L}}}{\tilde{Y}_{\mathcal{J}} S^{n+1}_{\mathcal{S},\mathcal{L}}} \right) H_{\mathcal{H},\mathcal{L}} \\ &= \sum_{\mathcal{J} \setminus \mathcal{S}} H_{\mathcal{H},\mathcal{L}} - \left(\sum_{\mathcal{J} \setminus \mathcal{S}} \frac{Y_{\mathcal{J}}}{\tilde{Y}_{\mathcal{J}}} H_{\mathcal{H},\mathcal{L}} \right) \frac{S^n_{\mathcal{S},\mathcal{L}}}{S^{n+1}_{\mathcal{S},\mathcal{L}}} \\ &= \partial^+_{S_{\mathcal{S},\mathcal{L}}} D(Y|S^n H) - \partial^-_{S_{\mathcal{S},\mathcal{L}}} D(Y|S^n H) \frac{S^n_{\mathcal{S},\mathcal{L}}}{S^{n+1}_{\mathcal{S},\mathcal{L}}} = 0, \end{aligned}$$

which easily implies $S^{n+1}_{\mathcal{S},\mathcal{L}} = S^n_{\mathcal{S},\mathcal{L}} \times \frac{\partial^-_{S_{\mathcal{S},\mathcal{L}}} D(Y|S^n H)}{\partial^+_{S_{\mathcal{S},\mathcal{L}}} D(Y|S^n H)}$, the desired multiplicative updates. □

The proof of lemma 14 is based on bounding $D_{KL}(Y|SH)$ by above using an auxiliary function, due to the convexity of d_{KL}. The result when Y, S and H are matrices is well explained in [176, 100]. From the intensity equations in section 3.2, the intensity tensor $\overline{\lambda}$ is a combination of sums and products of the tensors μ, S, J, B, b and $\overline{\phi}$, which makes lemma 14 suitable for the estimation of these parameters.

Unfortunately, the cost function (3.3) is not convex on the ensemble of tensors, which means that we cannot expect to retrieve the global minimum of $D_{KL}(Y|\overline{\lambda})$, i.e., the global maximum of the Hawkes likelihood. Nevertheless, it is convex (due to the convexity of the Kullback-Leibler divergence) on each tensor, given that the other is fixed. So, estimating each tensor given the rest fixed in a cyclic way produces nonincreasing values for Eqn. (3.3), as in [176, 100], thus converging to a local maximum of the approximated log-likelihood.

When $\delta \to 0$, the Riemann sums converge to their respective integrals, and minimizing the cost function in Eqn. (3.3) becomes equivalent to maximizing the likelihood of X.

As all information diffusion models of our Hawkes-based framework can be estimated using the same techniques based on lemmas 13 and 14, we have thus created an unified information dissemination framework using Hawkes processes.

Similarly to nonnegative matrix factorization (NMF) problems [193, 100], the multiplicative updates in lemma 14 can be sometimes written in a concise matrix form. We give next three examples of such cases: the models of subsections 3.2.1, 3.2.4 and 3.2.3.

3.3.1 Estimation of model in subsection 3.2.1

In order to proceed to the estimation of the Hawkes parameters J, B and μ, one needs first to handle the user-user interaction with care: due to the overwhelming number of user-user interaction

parameters $J_{i,j}$ in real-life social networks (where we have millions or even billions of users), we factorize J into FG, such that $F \in \mathcal{M}_{N\times d}(\mathbb{R}_+)$ is a $N \times d$ matrix and $G \in \mathcal{M}_{d\times N}(\mathbb{R}_+)$ is a $d \times N$ matrix, with $d \ll N$. This method is similar to clustering our social network influence graph into different communities (see [182]).

One can also notice that by performing a dimensionality reduction $J = FG$ during the estimation, we not only estimate the influence that users have over one another but we also acquired information on the communities of the underlying social network, since we were able to factorize the hidden influence graph J.

This is a very difficult problem, since the cyclic multiplicative updates destroy this relationship, and the only other way to satisfy the constraint in Eqn. (3.1) is to estimate each coordinate separately. Since $A_{i,j} \in \{0, 1\}$, we can circumvent this problem using a convex relaxation[7] of this constraint of the form[8] $\eta\langle 1 - A, FG\rangle$ and $\eta \geq 0$ a penalization parameter.

We have the following penalization $\eta\langle 1 - A, FG\rangle$, with derivatives

$$\begin{aligned}\nabla_F \eta\langle 1 - A, FG\rangle &= \eta(1 - A)G^T \text{ and}\\ \nabla_G \eta\langle 1 - A, FG\rangle &= \eta F^T(1 - A).\end{aligned} \tag{3.6}$$

Unfortunately, since F and G act as a product, there is a potential identifiability issue of the form $FG = F\mathcal{P}\mathcal{P}^{-1}G = \tilde{F}\tilde{G}$ where $\mathcal{P}$ is any scaled permutation and the pair $\tilde{F} = F\mathcal{P}$, $\tilde{G} = \mathcal{P}^{-1}G$ is also a valid factorization of J (see [192, 227, 245]). We deal with this issue normalizing the rows of G to sum to 1 (see [192, 245]). This normalization step involves the resolution of a nonlinear system for each row of G to find the associated Lagrange multipliers.

Our constraint thus becomes $G1 = 1$, for which the Karush-Kuhn-Tucker (KKT) conditions are written in matrix form as $\overline{\eta}_G = \sum_{i=1}^{d} \eta_{G,i} e_i^T 1$, with $(e_i)_{i\in\{1,\cdots,d\}}$ the standard basis vectors and $\eta_{G,i} \in \mathbb{R}$ the Lagrange multipliers solution of the nonlinear equation $G1 = 1$ after the update[9].

Let us recall that in this particular model we have the Hawkes parameters $J = FG, B$ and μ. In this particular model, one can further simplify the multiplicative updates given by lemma 14, using the structure of the intensity, as presented in [212].

We can redefine, with an abuse of language, the $NK \times T$ matrices $Y, \overline{\lambda}, \overline{\phi}$ and $\overline{\mu}$ as

$$\begin{aligned}Y_{i+(k-1)N,t} &= \frac{dX^{i,k}_{(t-1)\delta}}{\delta}\\ \overline{\lambda}_{i+(k-1)N,t} &= \lambda^{i,k}_{(t-1)\delta}\\ \overline{\phi}_{i+(k-1)N,t} &= (\phi * dX^{i,k})_{(t-1)\delta}\\ \overline{\mu}_{i+(k-1)N,t} &= \mu^{i,k},\end{aligned}$$

i.e., these matrices are the transposition of the mode-3 matricizations [170] of the $N \times K \times T$ dimensional tensors Y, $\overline{\lambda}$ and $\overline{\phi}$.

Let us also define the $N \times T$ matrices Y^k, $\overline{\lambda}^k$, $\overline{\phi}^k$ and $\overline{\mu}^k$ such that

$$Y^k_{i,t} = Y^{i,k}_t, \qquad \overline{\lambda}^k_{i,t} = \overline{\lambda}^{i,k}_t, \qquad \overline{\phi}^k_{i,t} = \overline{\phi}^{i,k}_t, \qquad \overline{\mu}^k_{i,t} = \mu^{i,k},$$

7. We use here a L^1 convex relaxation, similar to the LASSO procedure [284, 139]. This implies that we introduce a certain degree of sparsity on the influence graph J. Other kinds of convex relaxation functions can be used, such as a L^2 relaxation of the form $\eta||(1 - A) - FG||^2$.
8. From now on we denote by 1 any vector of matrix with entries equal to 1. The dimension of 1 will be clear in the context.
9. The same reasoning is applied to the matrices B, b and S defined in section 3.2, i.e., $\sum_k B_{c,k} = 1$, $\sum_k b_{i,k} = 1$ and $\sum_n S_{m,n} = 1$.

the $dK \times T$ matrix ρ and the $d \times T$ matrices ρ^k and $\overline{\rho}^k$ such that

$$\rho = (\mathbb{I} \otimes G)\overline{\phi}, \qquad \rho^k_{i,t} = \sum_j G_{i,j}\overline{\phi}^k_{j,t} \quad \text{and} \quad \overline{\rho}^k_{i,t} = \sum_{k'} B_{k',k}\rho^{k'}.$$

The $NK \times T$ intensity matrix $\overline{\lambda}$ can thus be written as (see [212])

$$\overline{\lambda} = \mu + (B^T \otimes J)\overline{\phi},$$

where $\otimes$ is the Kronecker product [156].

Following [212], we derive matrix-based multiplicative updates for the Hawkes paremeters F, G, B and μ.

Lemma 15. *We have the following multiplicative updates for F:*

$$F \leftarrow F \odot \frac{\sum_{k=1}^K \left([\frac{Y^k}{\overline{\lambda}^k}](\overline{\rho}^k)^T\right)}{\sum_{k=1}^K 1(\overline{\rho}^k)^T + \eta_F(1-A)G^T}, \tag{3.7}$$

where $\eta_F(1-A)G^T$, with $\eta_F \geq 0$, is a convex penalization term responsible for the constraint in Eqn. (3.1).

Proof. First of all, we have that $(B^T \otimes FG) = (B^T \otimes F)(\mathbb{I} \otimes G)$, thus $\overline{\lambda} = \overline{\mu} + (B^T \otimes F)(\mathbb{I} \otimes G)\overline{\phi}$.

Let F_i be the rows of F and ρ_t be the columns of ρ, with ρ^k_t the columns of the submatrices ρ^k. Then

$$\big((B^T \otimes F)\rho\big)_{i+(k-1)N,t} = \sum_{k'=1}^K B^T_{k,k'}\langle F_i, \rho^{k'}_t\rangle = \langle F_i, \sum_{k'=1}^K B_{k',k}\rho^{k'}_t\rangle = (F\overline{\rho}^k)_{it}.$$

Hence

$$\begin{aligned} D_{KL}(Y|\overline{\lambda}) &= \sum_{j,t} d_{KL}(Y_{jt}|\overline{\lambda}_{jt}) = \sum_t \sum_i \sum_k d_{KL}(Y_{i+(k-1)N,t}|\overline{\lambda}_{i+(k-1)N,t}) \\ &= \sum_k \left(\sum_t \sum_i d_{KL}(Y^k_{i,t}|\mu^{i,k} + (F\overline{\rho}^k)_{i,t}) \right) = \sum_k D_{KL}(Y^k|\overline{\mu}^k + F\overline{\rho}^k) \\ &= D^F_{KL}(F). \end{aligned}$$

One can see that D^F_{KL} is a sum of cost functions involving matrix products, hence we can use the well known nonnegative matrix factorization techniques [100, 193] to derive multiplicative updates for F. We have thus the following multiplicative update rule

$$F \leftarrow F \odot \frac{\sum_{k=1}^K \left([\frac{Y^k}{\overline{\lambda}^k}](\overline{\rho}^k)^T\right)}{\sum_{k=1}^K 1(\overline{\rho}^k)^T},$$

Since the penalization term $\eta_F(1-A)G^T$ has all its entries nonnegative, it is added to the denominator of the NMF updates, as in [100]. Following [100], we can rewrite the multiplicative updates with the linear penalization as Eqn. (3.7). □

Define the $N \times T$ auxiliary matrices $\overline{\Phi}^k$ such that

$$\overline{\Phi}^k = \sum_{k'} B_{k'k}\overline{\phi}^{k'},$$

i.e., $\overline{\lambda}^k = \overline{\mu}^k + FG\overline{\Phi}^k$.

Lemma 16. *We have the following multiplicative updates for G:*

$$G \leftarrow G \odot \frac{\sum_{k=1}^{K} F^T\left([\frac{Y^k}{\overline{\lambda}^k}](\overline{\Phi}^k)^T\right)}{\sum_{k=1}^{K} F^T 1(\overline{\Phi}^k)^T + \eta_G F^T(1-A) + \overline{\eta}_G}, \tag{3.8}$$

where $\overline{\eta}_G$ is a $d \times N$ matrix composed by Lagrange multipliers solution of the nonlinear equation $G1 = 1$ and $\eta_G F^T(1-A)$, with $\eta_G \geq 0$, is responsible for the constraint in Eqn. (3.1).

Proof. Firstly, we have that

$$\begin{aligned} D_{KL}(Y|\overline{\lambda}) &= \sum_{j,t} d_{KL}(Y_{jt}|\overline{\lambda}_{jt}) = \sum_{i,t,k} d_{KL}(Y^k_{i,t}|\mu^{i,k} + \langle \sum_{k'=1}^{K} B_{k'k} F_i, G\overline{\phi}^{k'}_t \rangle) \\ &= \sum_k \left(\sum_{i,t} d_{KL}(Y^k_{it}|\left(\overline{\mu}^k + FG\overline{\Phi}^k\right)_{i,t}) \right) = \sum_k D_{KL}(Y^k|\overline{\mu}^k + FG\overline{\Phi}^k) \\ &= D^G_{KL}(G). \end{aligned}$$

Using the same arguments as with F, we have the update rule for G given by Eqn. (3.8). □

Define the $K \times T$ matrices $\overline{Y}^i$, $\overline{\mu}^i$, $\overline{\lambda}^i$ and ζ^i such that

$$\begin{aligned} \overline{Y}^i_{k,t} &= Y_{i+(k-1)N,t} = Y^{i,k}_t \\ \overline{\mu}^i_{k,t} &= \overline{\mu}_{i+(k-1)N,t} = \mu^{i,k} \\ \zeta^i_{k,t} &= \left(J\overline{\phi}^k\right)_{i,t} = \sum_j J_{i,j}\overline{\phi}^k_{j,t} \\ \overline{\lambda}^i_{k,t} &= \overline{\lambda}_{i+(k-1)N,t} = \mu^{i,k} + \left(B^T\zeta^i\right)_{k,t}. \end{aligned}$$

Lemma 17. *We have the following multiplicative updates for B:*

$$B \leftarrow B \odot \frac{\sum_{i=1}^{N} \zeta^i [\frac{(\overline{Y}^i)^T}{(\overline{\lambda}^i)^T}]}{\sum_{i=1}^{N} \zeta^i 1 + \overline{\eta}_B}, \tag{3.9}$$

where $\overline{\eta}_B$ is a matrix composed by the Lagrange multipliers solution of the nonlinear equation $B1 = 1$.

Proof. Firstly, we have

$$\begin{aligned} D(Y|\overline{\lambda}) &= \sum_{j,t} d_{KL}(Y_{jt}|\overline{\lambda}_{jt}) = \sum_{i,t,k} d_{KL}(\overline{Y}^i_{k,t}|\mu^{i,k} + \sum_{k'} B^T_{kk'}\zeta^i_{k',t}) \\ &= \sum_i \left(\sum_{i,t} d_{KL}(\overline{Y}^i_{k,t}|\left(\underline{\mu}^i + B^T\zeta^i\right)_{k,t}) \right) = \sum_i D_{KL}(\overline{Y}^i|\overline{\mu}^i + B^T\zeta^i) = D^B_{KL}(B^T). \end{aligned}$$

By the same principle as in the estimation of F and G, the updates for B are given by Eqn. (3.9). □

Lemma 18. *Let v be the vectorization operation on matrices. We have the multiplicative updates for μ:*

$$v(\mu) = v(\mu) \odot \frac{[\frac{Y}{\overline{\lambda}}]1}{\langle 1, 1\rangle} = v(\mu) \odot \frac{[\frac{Y}{\overline{\lambda}}]1}{T}. \tag{3.10}$$

Proof. By the same token, it is easy to see that

$$\begin{aligned} D(Y|\overline{\lambda}) &= \sum_{j,t} d_{KL}(Y_{jt}|\big(v(\mu)1 + (B^T \otimes J)\overline{\phi}\big)_{jt}) \\ &= \sum_{j,t} d_{KL}(Y_{jt}|\big(v(\mu)1\big)_{jt} + \big((B^T \otimes J)\overline{\phi}\big)_{jt}) = D^{\mu}_{KL}(Y|\mu), \end{aligned}$$

giving us the multiplicative updates in Eqn. (3.10). □

3.3.2 Estimation of model in subsection 3.2.4

When using the simplest model for multiple social networks, if one assumes that[10] $\phi^n = \phi$, i.e., every social networks has the same temporal kernel for the past influences, one may write the intensity for this model in a concise way.

Let us recall that one needs first to handle the user-user interaction with care, and perform the dimensionality reduction $J = FG$, as in subsection 3.3.1. One must then introduce the convex penalizations given by Eqn. (3.6) in the denominator of the multiplicative updates given by lemma 14.

Let us also recall the intensity $N \times K \times M \times T$ tensor $\overline{\lambda}_{i,k,n,t}$ which represents the intensity of user i on broadcasting a message of topic k on social network n at time t. We also have the convolution $N \times K \times M \times T$ tensor $\overline{\phi}$, defined as

$$\overline{\phi}_{i,k,n,t} = (\phi * dX^{i,k,n})_t = \int_0^{t-} \phi(t-s) dX_s^{i,k,n} ds,$$

and the intrinsic Poissonian rate $N \times K \times M \times T$ tensor $\overline{\mu}$, defined as

$$\overline{\mu}_{i,k,n,t} = \mu^{i,k,n}.$$

The estimation for this model is hence quite straightforward, since the intensity tensor $\overline{\lambda}$ can be written as

$$\overline{\lambda} = \overline{\mu} + \overline{\phi} \times_1 J \times_2 B^T \times_3 S^T = \overline{\mu} + \overline{\phi} \times_1 G \times_1 F \times_2 B^T \times_3 S^T,$$

where $\times_l$ is the mode-l product between a tensor and a matrix [170, 66]. This means that the tensor $\overline{\lambda} - \overline{\mu}$ can be decomposed in a Tucker decomposition [170] with a nonnegative core tensor $\overline{\phi}$ and nonnegative matrices J, B^T and S^T.

Since by lemma 13 the maximum likelihood estimation is equivalent to a cost function using the Kullback-Leibler divergence, the nonnegative Tucker decomposition techniques satisfy lemma 14. They can be represented explicitly in a concise form using mode products and mode matricizations of tensors. The basic updates for nonnegative Tensor decomposition methods using the Kullback-Leibler divergence cost function can be found in [170].

3.3.3 Estimation of model in subsection 3.2.3

We concentrate in this subsection on the "fuzzy" diffusion model with user-user and topic-topic interactions in a single social network. The estimation procedure in "fuzzy" diffusions follow the same ideas as in the preceding subsections, with a minor difference: one also needs to estimate the

10. Actually, the only necessary hypothesis is that the temporal kernel in the intensity for $\lambda_t^{i,k,n}$ cannot depend on i, k or n.

topic model parameters. The topic model parameters are estimated during the Hawkes parameters estimation phase, and are influenced by the Hawkes process itself. At the same time, the topic model parameters also influence the Hawkes parameters estimation through the topic random variables Z.

The estimation step is thus performed in several steps in an iterative fashion: 1) estimate the topic model parameters, with the Hawkes parameters F, G, B, b and μ fixed. 2) using the conditional log-likelihood of X given Z and lemma 13, estimate the Hawkes parameters fixed in a cyclical way, 3) Repeat steps 1) and 2) until convergence.

We perform here only the estimation procedure for the Hawkes parameters, given the empirical topic proportions Z fixed. The modified estimation procedure for the topic model parameters is performed in appendix C, for the latent dirichlet allocation [34] and the author-topic [265] topic models.

Let us recall that one needs first to handle the user-user interaction with care and perform the dimensionality reduction $J = FG$, as in subsection 3.3.1. One must then introduce the convex penalizations given by Eqn. 3.6 in the denominator of the multiplicative updates given by lemma 14.

When estimating the Hawkes parameters in "fuzzy" diffusion models, one must fix the topic model random variables Z and maximize the conditional log-likelihood of X given Z. For the model in subsection 3.2.3, we have that this conditional log-likelihood is given by

$$\mathcal{L}(X|Z) = \log\left(\prod_{0\leq t_n\leq\tau}\lambda^{i_n}_{t_n}\right) - \sum_i\int_0^\tau \lambda_t^{i,m}dt = \sum_i\left(\int_0^\tau \log\lambda_t^i dX_t^i - \int_0^\tau\lambda_t^i dt\right).$$

As already proved in lemma 13, maximizing the Riemann-sum approximation of this conditional log-likelihood is equivalent to minimizing the cost function

$$D_{KL}(Y|\overline{\lambda}) = \sum_{i,t} d_{KL}(Y_t^i|\mu^i + \sum_k b_{i,k}\sum_j\sum_c\sum_l F_{i,l}G_{l,j}\overline{\phi}_t^{j,c}B_{c,k}). \tag{3.11}$$

We thus use again lemma 14 to derive concise multiplicative updates for F, G, B, b and μ, following the same ideas as in subsection 3.3.1:

Lemma 19. *Define the $dK \times T$ matrix $\rho = (\mathbb{I}\otimes G)\overline{\phi}$ and the $d\times T$ matrices ρ^k and $\overline{\rho}^k$ such that*

$$\rho^k_{i,t} = \rho_{i+(k-1)N,t} \quad \textit{and} \quad \overline{\rho}^k_{i,t} = \sum_{k'=1}^K B_{k',k}\rho^{k'}_{i,t},$$

and define the T row vectors $Y^n = (Y_1^n, Y_2^n, \cdots, Y_T^n)$ and $\overline{\lambda}^n = (\overline{\lambda}_1^n, \overline{\lambda}_2^n, \cdots, \overline{\lambda}_T^n)$.

Let F^n be the d row vector $F^n = (F_{n,1}, \cdots, F_{n,d})$, i.e., the n^{th} row of the matrix F, and let $(1-A)^n$ be the n^{th} row of the matrix $(1-A)$.

We have the following multiplicative updates for F^n:

$$F^n \leftarrow F^n \odot \frac{\sum_{k=1}^K b_{n,k}\left(\left[\frac{Y^n}{\overline{\lambda}^n}\right](\overline{\rho}^k)^T\right)}{\sum_{k=1}^K b_{n,k}1(\overline{\rho}^k)^T + \eta_F(1-A)^nG^T},$$

where $\eta_F(1-A)^nG^T$, with $\eta_F \geq 0$, is a convex penalization term responsible for the constraint in Eqn. (3.1)

Proof. We have by lemma 14 applied to the cost function (3.11) that

$$\begin{aligned}\partial_{F_{n,m}} D_{KL}(Y|\overline{\lambda}) &= \sum_t \left(1 - \frac{Y_t^n}{\overline{\lambda}_t^n}\right) \sum_k b_{n,k} (G\overline{\phi}_t B)_{m,k} \\ &= \sum_k \left(\sum_t b_{n,k} \left(1 - \frac{Y_t^n}{\overline{\lambda}_t^n}\right) (G\overline{\phi}_t B)_{m,k} \right).\end{aligned}$$

The conclusion follows easily by rearranging the positive and negative terms. □

We can apply the same principle to G, B, μ and b, yielding:

Lemma 20. *Define by abuse of notation the $N \times T$ matrices Y and $\overline{\lambda}$ such that*

$$Y_{i,t} = Y_t^i = \frac{dX_t^i}{\delta} \quad \text{and} \quad \overline{\lambda}_{i,t} = \overline{\lambda}_t^i.$$

Define also the $N \times d$ matrices F^k and the $N \times T$ matrices $\overline{\Phi}^k$ such that

$$F_{i,l}^k = F_{i,l} b_{i,k} \quad \text{and} \quad \overline{\Phi}_{i,t}^k = \sum_{k'} B_{k'k} \overline{\phi}_t^{i,k'}.$$

We have the following multiplicative updates for G:

$$G \leftarrow G \odot \frac{\sum_{k=1}^K (F^k)^T \left([\frac{Y}{\overline{\lambda}}] (\overline{\Phi}^k)^T \right)}{\sum_{k=1}^K (F^k)^T 1 (\overline{\Phi}^k)^T + \eta_G F^T (1 - A) + \overline{\eta}_G},$$

where $\overline{\eta}_G$ is a $d \times N$ matrix composed by Lagrange multipliers solution of the nonlinear equation $G1 = 1$ and $\eta_G F^T (1 - A)$, with $\eta_G \geq 0$, is a convex penalization term responsible for the constraint in Eqn. (3.1).

Lemma 21. *Let $Y^i = (Y_1^i, \cdots, Y_T^i)$ and $\overline{\lambda}^i = (\overline{\lambda}_1^i, \cdots, \overline{\lambda}_T^i)$ be T column vectors and B^k be the k^{th} column of B.*

Define the $K \times T$ matrices ζ^i such that

$$\zeta_{k,t}^i = \sum_j J_{ij} \overline{\phi}_t^{j,k}.$$

We have the following multiplicative updates for B^k:

$$B^k \leftarrow B^k \odot \frac{\sum_{i=1}^N b_{i,k} \zeta^i [\frac{Y^i}{\overline{\lambda}^i}]}{\sum_{i=1}^N b_{i,k} \zeta^i 1 + \overline{\eta}_B^k},$$

where $\overline{\eta}_B^k$ is the k^{th} column of the matrix $\overline{\eta}_B$, composed by the Lagrange multipliers solution of the nonlinear equation $B1 = 1$.

Lemma 22. *Let $Y^n = (Y_1^n, \cdots, Y_T^n)$ and $\overline{\lambda}^n = (\overline{\lambda}_1^n, \cdots, \overline{\lambda}_T^n)$ be T row vectors and b^n be the n^{th} row of b.*

Define the $K \times T$ matrices Ψ^n such that

$$\Psi_{k,t}^n = \sum_{l,c} J_{n,l} \overline{\phi}_t^{l,c} B_{c,k} = (J\overline{\phi}_t B)_{n,k}.$$

We have the following multiplicative updates for b^n:

$$b^n \leftarrow b^n \odot \frac{[\frac{Y^n}{\overline{\lambda}^n}](\Psi^n)^T}{1(\Psi^n)^T + \overline{\eta}_b^n},$$

where $\overline{\eta}_b^n$ is the n^{th} row of the matrix $\overline{\eta}_b$, composed by the Lagrange multipliers solution of the nonlinear equation $b1 = 1$.

Lemma 23. *Define by abuse of notation the $N \times T$ matrices Y and $\overline{\lambda}$ such that*

$$Y_{i,t} = Y_t^i = \frac{dX_t^i}{\delta} \quad \text{and} \quad \overline{\lambda}_{i,t} = \overline{\lambda}_t^i.$$

We have the following multiplicative updates for μ:

$$\mu = \mu \odot \frac{[\frac{Y}{\overline{\lambda}}]1}{1^T 1} = \mu \odot \frac{[\frac{Y}{\overline{\lambda}}]1}{T}.$$

Remark: We derive in [211] an approximated estimation procedure based on weighted nonnegative matrix factorization techniques [36] for F, G, B, μ and b, using the convexity of d_{KL} and the logarithm in the maximum likelihood of X. This procedure maximizes a function that bounds by below the conditional log-likelihood of X.

Algorithm 2 - Hawkes estimation procedure

Input: jumps dX, step size δ, temporal kernels $(\phi^m)_{m\in\{1,\cdots,M\}}$
Discretize $[0, \tau]$ into T bins of size δ
Calculate normalized jumps $Y = \frac{dX}{\delta}$, convolution tensors $\overline{\phi}$ and discretized intensities $\overline{\lambda}$
Initialize Hawkes matrices set $\mathcal{X}$ *(for example, in a user-user topic-topic model with predefined topics $\mathcal{X} = \{F, G, B, S, \mu\}$)*
while Matrices in $\mathcal{X}$ have not converged **do**
    if In a "fuzzy" diffusion model **then**
        With all Hawkes matrices $\mathcal{X}$ fixed, run round of topic model estimation as dictated by appendix C
    end if
    for matrix x in $\mathcal{X}$ **do**
        With all other matrices fixed *(and topic model parameters as well, if in a "fuzzy" diffusion model)*, update x as

$$x^{n+1} \leftarrow x^n \odot \frac{\nabla_x^- D_{KL}(Y|\overline{\lambda})_{|x^n}}{\nabla_x^+ D_{KL}(Y|\overline{\lambda})_{|x^n}},$$

    end for
end while
Output: Hawkes matrices $\mathcal{X}$ and topic model parameters

3.4 Additional remarks

This section is concerned with further explanations and extensions of the Hawkes information diffusion framework developed in this chapter. The discussion and explanations regard solely the

Hawkes process, which means that it is not concerned with aspects on topic models. They are discussed at length in appendix C.

We start by calculating, as an illustrative example, the complexity of the multiplicative updates given by lemma 14 applied to the estimation of the model in subsection 3.2.1, which were derived in matrix form in subsection 3.3.1.

3.4.1 Complexity of the estimation procedure in subsection 3.3.1

As shown in subsection 3.3.1, the nonnegative tensor decomposition updates for the model in subsection 3.2.1 can actually be written in matrix form, which in this case give rise to modified nonnegative matrix factorization updates.

Nonnegative matrix factorization techniques are multiplicative updates, using only entrywise operations and matrix products[11], which are fast and can be performed in a distributed fashion very easily. Hence, at each step of the cyclic estimation procedure, we have the following complexity for the updates, written in terms of the number of users N, the number of topics K, the factorization dimension d and the number of time discretization steps T:

Hence, at each step of the cyclic estimation procedure, we have the following complexity for the updates:

3.4.1.1 Complexity for F

(1) First we need to calculate $\rho = (\mathbb{I}_{K\times K} \otimes G)\overline{\phi}$, which is of complexity $\mathcal{O}(dNKT)$ if the structure of the Kronecker product is exploited.

(2) Then, following Eqn. (3.7), we have that the complexity for the numerator update of F is $K \times \mathcal{O}(dNT) = \mathcal{O}(dKNT)$, while for the denominator is $\mathcal{O}(dKNT) + K \times \mathcal{O}(dN^2) = \mathcal{O}(dKNT + dN^2)$ due to the penalization term (one can see that the real complexity of this term is much lower since A is in practice a sparse matrix).

Thus, the complexity of updating F is $\mathcal{O}(dKNT + dKN^2)$.

3.4.1.2 Complexity for G

(1) First we need to calculate $\overline{\Phi}^k$ for all k, which is of complexity $\mathcal{O}(NK^2T)$.

(2) Then, following Eqn. (3.8), we have that the complexity for the numerator update of G is $K \times \mathcal{O}(dNT) = \mathcal{O}(dKNT)$ if done in the proper order (there are two matrix products in the numerator this time), while for the denominator it is $\mathcal{O}(dKNT + dN^2)$ due to the penalization term.

Thus, the complexity of updating G is $\mathcal{O}(K^2NT + dKNT + dKN^2)$.

Remark: For the complexity of G and B, we also have to take into consideration the calculation of the Lagrange multipliers $\overline{\eta}_G$ and $\overline{\eta}_B$. These multipliers are calculated using convex optimization techniques[12], whose complexity is not greater than the complexity of the multiplicative updates for G or B.

11. We make use of the fact that the complexity of a product of two matrices of sizes $n \times m$ and $m \times t$ is at most $\mathcal{O}(nmt)$.
12. Since we need to find the zero of the function $h(\eta) = \frac{1}{a+\eta}$.

3.4.1.3 Complexity for B

(1) First we need to calculate ζ^i for all $i \in V$, which is of complexity $\mathcal{O}(dNKT)$ if done using the fact that $J = FG$.

(2) Then, following Eqn. (3.9), we have that the complexity for the numerator and denominator updates of B is $N \times \mathcal{O}(K^2T) = \mathcal{O}(K^2NT)$.

Thus, the complexity of updating B is $\mathcal{O}(dNKT + K^2NT)$.

3.4.1.4 Complexity for μ

Following Eqn. (3.10), we have that the complexity for μ is simply $\mathcal{O}(dKNT)$, given by the calculation of λ, which is the lowest complexity of all updates.

3.4.1.5 Total complexity of the updates

We can clearly see that the calculation of F is of complexity $\mathcal{O}(dKNT + dKN^2)$, that of G is of complexity $\mathcal{O}(K^2NT + dKNT + dKN^2)$, that of B is of complexity $\mathcal{O}(dKNT + K^2NT)$ and that of μ is of complexity $\mathcal{O}(dKNT)$. Hence the complexity to update at each cyclical estimation step is

$$\mathcal{O}(K^2NT + dKNT + dKN^2).$$

Remark: It is worthy mentioning that the quadratic complexity of $\mathcal{O}(dKN^2)$ due to the penalization terms can be reduced in several ways: first, the term $(1 - A)$ appearing in the matrix products $(1 - A)G^T$ and $F^T(1 - A)$ is normally sparse for real-life social networks. Second, one may use surrogates for the matrix $(1 - A)$, for example dividing the graph G into communities beforehand and using the communities as representatives for the users. Third, one may use simply the initial implementation on F and G during the estimation steps, since a strictly positive entry for F and G continues to be strictly positive during the entire procedure. Fourth, since the convex penalizations are matrix products, one can calculate them in a distributed fashion, reducing drastically the complexity.

Thus, if we assume $K \ll N$, $K \ll T$ and $d \ll N$ and set aside the quadratic complexity from the penalization terms in F and G, we achieve the following complexity

$$\mathcal{O}(dKNT)$$

which is linear in every parameter, and dictated by N.

3.4.1.6 Complexity without the factorization $J = FG$

Following the same calculations as for the complexity of F using Eqn. (3.7), we get that the complexity for J is $K \times \mathcal{O}(N^2T) = \mathcal{O}(KN^2T)$.

By the same token, every time we factorize $J = FG$ to compute the other multiplicative updates for B and μ, we have to calculate $\overline{\lambda}$, which has a complexity of $\mathcal{O}(dKNT)$. If we cannot factorize J, the complexity becomes $\mathcal{O}(KN^2T)$, which is much larger than $\mathcal{O}(dKNT)$ since $d \ll N$.

This proves that the dimensionality reduction $J = FG$ is crucial to obtain a low complexity in the data.

3.4.2 Initial condition

One known setback in the nonnegative tensor factorization framework is the convergence to local minima of the cost function $D_{KL}(Y|\overline{\lambda})$, which means that a good initial condition is crucial for the estimation. There are results in the NMF literature that illustrate how to achieve a better estimation by constructing an improved initial condition (see [6, 42, 305]), nevertheless they do not work in our framework: our cost function is with respect to $D_{KL}(Y|\overline{\lambda})$ and the frameworks in [6, 42] do not apply if we consider finding good initial conditions for $J = FG$, B and μ *at the same time*. Moreover, we do not know the true value of $J = FG$, our only proxy is the adjacency matrix A, which is binary ($A_{i,j} \in \{0, 1\}$) and make it very hard to use the methods in [42, 305].

We use random initial conditions for B and μ, and we factorize A into $A = F_0 G_0$ using a standard NMF algorithm, with $F_0 \in \mathcal{M}_{N\times d}(\mathbb{R}^+)$ and $G_0 \in \mathcal{M}_{d\times N}(\mathbb{R}^+)$. We use then F_0 as the initial condition for F and G_0 as the initial condition for G.

That being said, different methods for the random initialization of the Hawkes parameters can be applied at one's desire, such as simulated annealing methods [171, 294], rank-by-rank (greedy) heuristics [294], multilayer techniques [65, 63], particle based and nature-inspired methods [164, 165, 166], etc.

3.4.3 Alternative estimation methods

The problem of nonnegative tensor factorization (or more specifically nonnegative matrix factorization) has been studied for a long time now, with a vast and varied research literature.

The NTD multiplicative updates used in this chapter are simply *one* of the existing methods for NTD estimation. The reasons for the use of multiplicative NTD updates are: they are easy to implement, can be implemented in a distributed fashion, have a low (even linear) complexity on the data, provide an easy way to introduce penalizations and constraints, and they provide a mathematically solid and unified estimation framework for the Hawkes-based information diffusion models.

Other methods wildly used are: projected gradient and alternate least-square algorithms [251, 62, 207], fixed-point alternating least-squares algorithms [64, 220], quasi-Newton algorithms [65, 315], multilayer techniques and hierarchical methods [250, 65, 61], etc. The reader has the excellent review of these methods in [63].

3.4.4 Extensions

This subsection is dedicated to the discussion of extensions of this information diffusion framework to accommodate different patterns on the data, and different behaviors of the model.

3.4.4.1 Nonparametric choice of the community structure parameter d

Our estimation method, which is based on maximum likelihood estimates of the point process X and nonnegative matrix factorization (NMF) techniques, requires the NMF parameter d. This parameter is intimately related to the community structure of the influence matrix J, as in [182], and is unfortunately ad-hoc, so it must be learned beforehand.

Tan and Févotte, however, derive an automatic way of finding the optimal d during the NMF updates in [277]. They do so by considering the NMF procedure for the β-divergence (for which the Kullback-Liebler divergence is a particular case) as a Bayesian estimation of an underlying probabilistic model.

One can thus introduce this automatic detection step into algorithm 2 in order to create a more data-driven estimation procedure for the Hawkes parameters.

3.4.4.2 Introduction of seasonality in the intrinsic intensity μ

It may be desirable to introduce periods in which people behave differently and thus broadcast messages differently; for example, users probably have a higher intrinsic rate during the lunch hour or the evening compared to the late night, since users are most certainly sleeping.

That being said, let us define nonoverlapping periods $\tau_n \in [0, \tau]$ such that $\tau_i \cap \tau_j = \emptyset$ and $\bigcup_n \tau_n = [0, \tau]$, following this practical example: let τ_1 be all the periods $[0, 6h]$ for every day in $[0, \tau]$, $\tau_2 = (6h, 12h]$, $\tau_2 = (12h, 18h]$ and $\tau_2 = (18h, 24h]$. We have divided thus $[0, \tau]$ into four regions, where each region is responsible for one quarter of every day in $[0, \tau]$.

Let 1^{τ_n} be the T vector such that $1_t^{\tau_n} = \mathbb{I}_{\{\delta(t-1) \in \tau_n\}}$ and μ_{τ_n} be the intrinsic rate associated with the period τ_n. Thus, $1 = \sum_n 1^{\tau_n}$ and we can apply our NTF procedure for each μ_{τ_n} separately, since

$$D_{KL}(Y|\overline{\lambda}) = D_{KL}(Y|\mu + SE) = D_{KL}(Y|\sum_n \mu_{\tau_n} + SE),$$

where SE accounts to the self-excited part of the intensity.

For example, for the estimation procedure of the model in subsection 3.2.3, we have the following updates for the periods τ_n:

$$\mu_{\tau_n} \leftarrow \mu_{\tau_n} \odot \frac{[\frac{Y}{\overline{\lambda}}]1^{\tau_n}}{\langle 1, 1^{\tau_n} \rangle}.$$

If dividing the time frame $[0, \tau]$ into different periods is not sufficient to account for temporal effects in the data and one must adopt a nonlinear behavior for the intrinsic rate μ (which is equivalent to consider the underlying Poisson process that generates the self-exciting cascades [141, 44] to be inhomogeneous), one can incorporate a nonparametric estimation of the intrinsic rate μ as in Lewis and Mohler [198].

3.4.4.3 Estimation of the temporal kernel

Another important part of the Hawkes modeling and estimation procedure is the shape of the temporal kernels $(\phi^m)_{m \in \{1, \cdots, M\}}$. The temporal kernels are responsible for the type of temporal interactions that broadcasts have when cascading through the social networks.

These interactions can appear in various forms, for example exponential temporal kernels imply short-range interactions and power-law temporal kernels imply long-range interactions. The parametric kernels introduce timescale parameters, which may be advantageous to retrieve from data, instead of being an input of the model. See subsection 3.2.4 for a more detailed discussion.

However, when estimating the temporal kernels ϕ^m, the convolution $\phi^m * dX$ must be recalculated[13] at each NTD cyclical update step, which increases the running time of the algorithm.

There are two alternatives to estimating the temporal kernels: parametric and nonparametric methods.

- Parametric kernels: they belong to basically two families of functions, exponential [245, 246, 310] and power-law [71] functions. They can be estimated with the maximization of the

13. When estimating exponential kernels, one could calculate $\phi^m * dX$ only up to a fixed lag, as in [310], which speeds up the algorithm. This is due to the fact that exponential kernels have light tails, which give more importance to the immediate past than the far away past.

Hawkes log-likelihood [15], with expectation-maximization methods [198, 221, 319], or with quadratic contrast minimization methods [18, 19].

Expectation-maximization methods are preferred over likelihood maximization methods, since maximum likelihood methods recur to numerical optimization algorithms and are costlier than their counterparts. Some expectation-maximization algorithms for exponential and power-law kernels are derived in appendix B, regarding the model in subsection 3.2.1.

Recently, however, Bacry *et al.* [18, 19] developed new concentration inequalities for quadratic contrast minimization in order to derive more data-driven, low-rank and sparse structures for information cascades.

- There are also attempts to derive nonparametric estimation of kernels for Hawkes processes, as in Lewis and Mohler [198], Al Dayri *et al.* in [80], Bacry and Muzy [17], an many more [262, 135, 16].

  Lewis and Mohler derive in [198] a kernel density estimation method and a maximum penalized likelihood estimation method, which results in Euler-Lagrange equations for the temporal kernel (and the intrinsic rate as well).

  Hansen *et al.* [135], Reynaud-Bouret and Schbath [262], and Reynaud-Bouret *et al.* [261] develop nonparametric estimation techniques for the temporal kernels in Hawkes processes using quadratic contrast minimization and oracle inequalities, with applications to neurobiology. It is worthy mentioning that the benefits of using contrast minimization is that it allows the estimation of nonlinear models (which can include negative temporal kernels) [135, 262] and the discovery of the "best" Hawkes process in accordance to the data (it may be that the assumption of data behaving as a multivariate Hawkes process is simply wrong).

  Bacry and Muzzy [17] and Bacry *et al.* [16] show that a Hawkes process can be characterized by its first and second order statistics, and derive nonparametric estimates for the temporal kernels based on a Wiener-Hopf system [237, 13].

  The problem with nonparametric kernel estimation is the high dimension of our Hawkes processes, i.e., $N \gg 1$. These methods have at least a quadratic complexity in N, which make them quite slower than the parametric alternatives, such as expectation-maximization methods, and impractical for real-life social networks.

3.4.4.4 Extension of dynamic/temporal networks

In many cases links on social networks are severed or acquired, which means that the social network in question may be a dynamic object [154], instead of a static one. Let us consider the model in subsection 3.2.1 and assume, without any loss of generality, that the topic-topic interaction matrix B remains static, for the sake of simplicity.

One has thus P increasing periods of time $(\tau_p)_{p\in\{1,\cdots,P\}}$ such that

- $\tau_p \subset [0,\tau]$, $\tau_{p'} \cap \tau_p = \emptyset$,
- $\bigcup_p \tau_p = [0,\tau]$ and
- $\sup \tau_p = \inf \tau_{p+1}$.

We also assume that the adjacency matrices for each time period τ_p, denoted by A^p, satisfy $A^p \neq A^{p'}$ if $p \neq p'$, i.e., each period of time represents a change on the underlying social structure. Let

- 1^{τ_p} be the $N \times T$ matrix such that $1^{\tau_p}_{i,t} = \mathbb{I}_{\{(t-1)\delta \geq \inf \tau_p\}}$,
- $p_t = \{p \in \{1, \cdots, P\} \mid t \in \tau_p\}$ be the unique index p such that $t \in \tau_p$,
- $\mathcal{P}_t = \bigcup_{s \leq t} p_s$ be all time period indices until time t and
- $j \overset{p}{\rightsquigarrow} i$ means that user j can influence user i on the time period τ_p.

The intensity in this model, as already studied before, is thus[14]

$$\begin{aligned}\lambda_t^{i,k} &= \mu^{i,k} + \sum_c B_{c,k} \int_0^{-t} \sum_{j \overset{p_s}{\rightsquigarrow} i} J^{p_s}_{i,j} \phi(t-s) dX_s^{j,c} \\ &= \mu^{i,k} + \sum_c B_{c,k} \sum_{p \in \mathcal{P}_t} \int_{\inf \tau_p}^{t \wedge \sup \tau_p} \sum_{j \overset{p}{\rightsquigarrow} i} J^{p}_{i,j} \phi(t-s) dX_s^{j,c},\end{aligned}$$

which in matrix form is given by (according to subsection 3.2.1)

$$\overline{\lambda}_t = \mu + \sum_{p \in \mathcal{P}_t} J^p \overline{\phi}^{p,t} B,$$

where the $N \times K$ matrices $\overline{\phi}^{p,t}$ satisfy

$$\overline{\phi}^{p,t}_{j,c} = \begin{cases} \int_{\inf \tau_p}^{(t-1)\delta \wedge \sup \tau_p} \phi(t-s) dX_s^{j,c} & \text{if } \ (t-1)\delta \in \tau_p \\ \\ 0 \ \text{ otherwise.} \end{cases}$$

By lemma 13, we have that our maximum likelihood estimation algorithm for F^p, G^p, B and μ is found by seeking the minimum of

$$D_{KL}(Y|\overline{\lambda}) = \sum_{i,k,t} d_{KL}(Y_t^{i,k} | \mu^{i,k} + \sum_{p \in P_{(t-1)\delta}} (F^p G^p \overline{\phi}^{p,t} B)_{i,k}).$$

Define the $N \times T$ matrices $Y^{p,k}$ and $\overline{\lambda}^{p,k}$, and the $d \times T$ matrices $\rho^{p,k}$ and $\overline{\rho}^{p,k}$, such that

$$\begin{aligned} Y^{p,k}_{i,t} &= Y^{i,k}_t . \mathbb{I}_{\{(t-1)\delta \geq \inf \tau_p\}}, \qquad \overline{\lambda}^{p,k}_{i,t} = \overline{\lambda}^{i,k}_t . \mathbb{I}_{\{(t-1)\delta \geq \inf \tau_p\}}, \\ \rho^{p,k}_{i,t} &= \sum_j G^p_{i,j} \overline{\phi}^{p,k}_{j,t} \quad \text{and} \quad \overline{\rho}^{p,k}_{i,t} = \sum_{k'=1}^{K} B_{k',k} \rho^{p,k'}_{i,t}. \end{aligned}$$

Using lemma 14 we have that the multiplicative updates for F take the form

$$F^p \leftarrow F^p \odot \frac{\sum_{k=1}^K [\frac{Y^{p,k}}{\overline{\lambda}^{p,k}}] (\overline{\rho}^{p,k})^T}{\sum_{k=1}^K 1^{\tau_p} (\overline{\rho}^{p,k})^T + \eta_F^p (1 - A^p)(G^p)^T},$$

with similar multiplicative updates for G^p. The multiplicative updates for B differ only in the auxiliary matrices, and those for μ remain unaltered.

14. One may see that this dynamical framework for J is similar to the convolutive nonnegative matrix factorization framework of [272, 308].

One may also be interested in finding the ensemble $(J^p)_{p\in\{1,\cdots,P\}}$ the smoothest as possible (if one sees the temporal function $p \mapsto J^p$ as the way the adjacency matrix J evolves through time). One may apply L^1 or L^2 regularization techniques during the estimation of $(F^p)_{p\in\{1,\cdots,P\}}$ and $(G^p)_{p\in\{1,\cdots,P\}}$.

The regularization procedure may occur in at least two different ways:

(i) The first way is to use a Tikhonov regularization g on the derivative of $p \mapsto J_p$, as $g(F,G) = \frac{\eta}{2}\sum_{0\leq p<P}||J^p - J^{p+1}||^2 = \frac{\eta}{2}\sum_{0\leq p<P}||F^pG^p - F^{p+1}G^{p+1}||^2$. We have derivatives[15]

$$\nabla_{F^p} g(F,G) = \eta(2J^p - J^{p-1} - J^{p+1})(G^p)^T = \eta(2F^pG^p - J^{p-1} - J^{p+1})(G^p)^T$$
$$\nabla_{G^p} g(F,G) = \eta(F^p)^T(2J^p - J^{p-1} - J^{p+1}) = \eta(F^p)^T(2F^pG^p - J^{p-1} - J^{p+1}),$$

with obviously $J^{-1} = 0$ and $J^{P+1} = 0$. It has positive part

$$\nabla^+_{F^p} g(F,G) = 2\eta J^p(G^p)^T = 2\eta F^pG^p(G^p)^T$$
$$\nabla^+_{G^p} g(F,G) = 2\eta(F^p)^T J^p = 2\eta(F^p)^T F^pG^p$$

and negative part

$$\nabla^-_{F^p} g(F,G) = \eta(J^{p+1} + J^{p-1})(G^p)^T$$
$$\nabla^-_{G^p} g(F,G) = \eta(F^p)^T(J^{p+1} + J^{p-1}).$$

It is thus easy to incorporate this step into the cyclical estimation algorithm 2 using lemma 14 and cost function $D_{KL}(Y|\overline{\lambda}) + g(F,G)$.

(ii) The second way is to first estimate the base matrix J^0 without regularization (which makes J^0 the "true"or "expected" value of J), and then at each time period τ_{p+1} we apply a regularization with respect to the past period τ_p, with $g(F^{p+1},G^{p+1}) = \frac{\eta^p}{2}||J^{p+1} - J^p||^2 = \frac{\eta^p}{2}||F^{p+1}G^{p+1} - J^p||^2$ as regularizing function. We have derivatives

$$\nabla_{F^{p+1}} g(F^{p+1},G^{p+1}) = \eta^{p+1}(F^{p+1}G^{p+1} - J^p)(G^{p+1})^T$$
$$\nabla_{G^{p+1}} g(F^{p+1},G^{p+1}) = \eta^{p+1}(F^{p+1})^T(F^{p+1}G^{p+1} - J^p),$$

which have positive part

$$\nabla^+_{F^{p+1}} g(F^{p+1},G^{p+1}) = \eta^{p+1}F^{p+1}G^{p+1}(G^{p+1})^T$$
$$\nabla^+_{G^{p+1}} g(F^{p+1},G^{p+1}) = \eta^{p+1}(F^{p+1})^T F^{p+1}G^{p+1}$$

and negative part

$$\nabla^-_{F^{p+1}} g(F^{p+1},G^{p+1}) = \eta^{p+1}J^p(G^{p+1})^T$$
$$\nabla^-_{G^{p+1}} g(F^{p+1},G^{p+1}) = \eta^{p+1}(F^{p+1})^T J^p.$$

It is thus easy to incorporate this step into the cyclical estimation algorithm 2 with lemma 14 by beginning with F^0 and G^0 (which do not require this regularization), and using the old values of F^p and G^p to estimate the new F^{p+1} and G^{p+1}.

Remark: One can see that this choice of regularization is equivalent (at least in principle) to assume that $J^{p+1} = J^p + \delta^p$, where δ^p is the derivative of $p \mapsto J^p$.

15. One can easily realize that the operation $p \to 2J^p - J^{p+1} - J^{p-1}$ is the numerical second derivative of the function $p \mapsto J^p$.

3.4.4.5 Nonparametric choice of K, number of topics, in "fuzzy" diffusion models

This part is the only extension regarding topic models, and is discussed here instead of appendix C because it also affects the Hawkes parameters.

Yang and Zha also comment in [310] that a nice extension of the topic would be to incorporate a nonparametric way of choosing the number of topics for the Hawkes model, i.e., the value of K. The natural mathematical framework for an unbounded (as yet to be discovered) number of topics is developed by Delattre *et al.* in [84], using infinite dimensional Hawkes processes.

The normalization $\sum_k B_{c,k} = 1$ allows one to easily satisfy the existence conditions in [84] of infinite dimensional Hawkes process for the model proposed, so if ones uses the infinite dimensional framework for Hawkes processes developed in [84] together with a topic model that accommodates a nonparametric selection of the number of topics, as in [281, 280], one can produce even more coherent and data-driven models and methods for information diffusion in the Hawkes framework.

3.5 Numerical examples

In this section we describe some numerical examples of this information diffusion framework, or more specifically, exemples of the model in subsection 3.2.1.

We have four different datasets, two simulated with the thinning algorithm[16] developed by Ogata in [241] and two real-life datasets:

- The first example is a synthetic dataset of a 2-clique uniformly random network with $N = 100$ (each complete clique having 50 nodes), $K = 10$ and an exponential temporal kernel, and concerns figures 3.1, 3.2, 3.3 and 3.4.

  We used $d = 51$ for our factorization $J = FG$, with a linear penalization (as in lemma 15) with constants $\eta_F = \eta_G = 10^3$. We did not use cross-validation techniques to find optimal penalization parameters η_F and η_G, since the algorithm is robust enough with respect to them.

  Figure 3.1 is the heatmap of $J = FG$, where the left heatmap is the estimated $J = FG$ and the right heatmap is the true value for J. One can clearly see that our algorithm retrieves quite well the structure behind the true J, i.e., two distinct cliques.

  Figure 3.2 is the heatmap of the squared difference of the true J and its estimation $\tilde{J}$, i.e., for each true entry $J_{i,j}$ and estimated entry $\tilde{J}_{i,j}$ we have plotted the differences $(J_{i,j} - \tilde{J}_{i,j})^2$ and $(J_{i,j} - \tilde{J}_{i,j})^2 / J_{i,j}^2$ (when $J_{i,j}$ is nonzero).

  Figure 3.3 refers to the squared difference of B and its estimation and figure 3.4 refers to the squared difference of the true μ and its estimation, as in figure 3.2.

- The second example is again a synthetic dataset, simulated for a 2-clique uniformly random network with $N = 20$ and $K = 1$ and exponential temporal kernel, and concerns figures 3.5 and 3.6.

  We compare our estimation choosing $d = 10$ with the estimation algorithm in [310] (with the obvisouly simplification of $K = 1$ and no language model), which models memes propagation in a social network using a Hawkes model similar to ours (identical to ours when $K = 1$),

16. The thinning algorithm simulates a standard Poisson process P_t with intensity $M > \sum_{i,k} \lambda_t^{i,k}$ for all $t \in [0, \tau]$ and selects from each jump of P_t the Hawkes jumps of $X_t^{i,k}$ with probability $\frac{\lambda_t^{i,k}}{M}$, or no jump at all with probability $\frac{M - \sum_{i,k} \lambda_t^{i,k}}{M}$.

but making use of an auxiliary language model for the memes labeling and not using the factorization $J = FG$ as in subsection 3.3.1; one can see that our algorithm (on the left of figure 3.5) outperforms the algorithm of [310] not only on the estimation[17] of μ, but also on the estimation of J, retrieving the community structure when the algorithm in [310] did not. Moreover, the algorithm of [310] needs an ad-hoc parameter ρ to control the sparsity of the network, which is not needed in our case.

- The third example is a Game of Thrones[18] (GOT) dataset with the dialogues of the pilot episode, with their respective timestamps and characters. We assumed that every character could influence all the others, and we estimated the characters hidden influence matrix J using $K = 1$, i.e., we are only concerned with the characters' influence on each other without any topic distinction. The heatmap of J is plotted on figure 3.7, and shows that our estimation algorithm indeed performs a community detection procedure, by dividing the influence graph into the two most famous families *Stark* and *Lannister*.

- The last example is a MemeTracker dataset, with different topics and world news for the $5,000$ most active websites from 4 million sites from March 2011 to February 2012[19]. We used the 5 most broadcasted memes, i.e., $K = 5$, leading to the websites influence graph in figure 3.8. This graph was plotted with the websites having the 10% largest outdegrees[20] and shows the influence of websites on one another. The thicker the edge lines, the larger the influence, and the larger the website's name, the larger the overall influence of the website (the sum of its influences).

3.6 Conclusion

We presented in this chapter a general framework to model information diffusion in social networks based on the theory of self-exciting point processes - linear multivariate Hawkes processes. Hawkes processes were already successfully introduced in a multitude of domains, such as neuroscience, finance, seismology, and even social sciences, and present themselves as a natural way to model information cascades in social networks.

The framework developed here exploits the real broadcasting times of users - a feature that comes with no mathematical overhead since we do so in the theory of point processes - which guarantees a more realistic view of the information diffusion cascades.

Our framework takes into consideration every possible type of influence between users and contents in social networks, under a variety of assumptions, which provides a deeper and much more general analysis of hidden influences in social networks.

This framework is also interesting for several other reasons: first, it allows one to use predefined topics (labeled data) and unknown topics (unlabeled data). The overhead of introducing topic

17. One can clearly see that our algorithm is able to detect the different sets of values for μ, although with a high variance. This is completely understandable, because a linear Hawkes process is equivalent to an Poisson cluster process (see [141]), where immigrants arrive following a Poisson process with rate μ. This means that the algorithm estimates a rate μ of a Poisson process, which is known to have (optimal) variance μ itself (see [92]), hence a larger rate implies a larger variance. Of course the estimation improves when $\tau \to \infty$, since for J fixed this is equivalent to a maximum likelihood estimator (MLE), which is consistent and asymptotically normal (see [240]); we used in this example a rather small τ due to performance reasons.
18. http://en.wikipedia.org/wiki/Game_of_Thrones.
19. Data available at http://snap.stanford.edu/netinf.
20. This means that we have chosen the 9^{th} decile of nodes regarding the distribution $(\sum_i J_{i,j})_{j \in V}$.

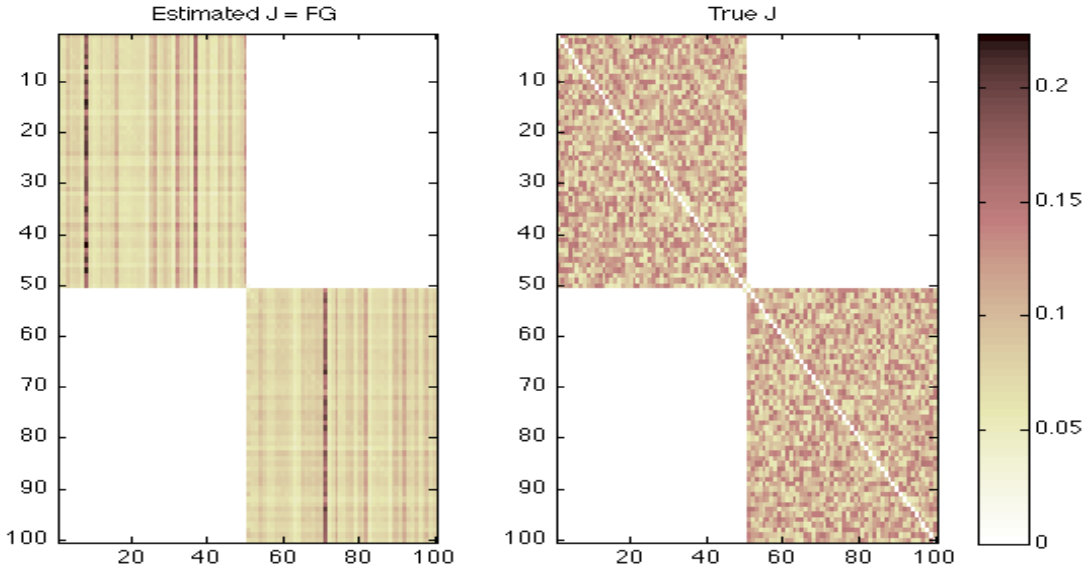

Figure 3.1: Heatmap of $J = FG$ for 2-clique network of 100 nodes.

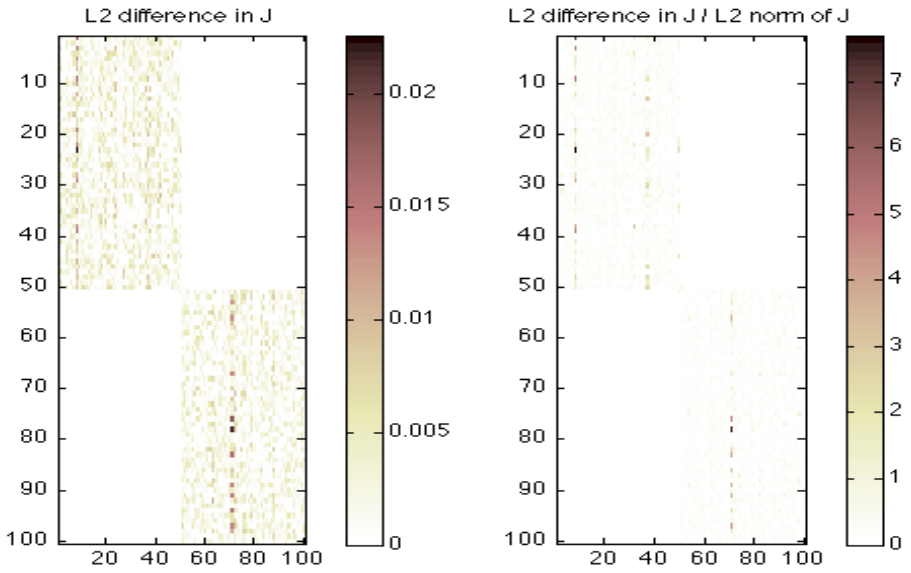

Figure 3.2: Heatmap of L^2 differences (absolute and relative) between entries of true J and estimated J.

models into the Hawkes models is minimal and allows a much more data-driven way of discovering the hidden influences on social networks, for which modified collapsed Gibbs sampling and variational Bayes techniques are derived; moreover, the generality of these topic models also simplifies the extension of our framework to any kind of data modeling, such as hierarchical topic models, semi-supervised topic models, nonparametric topic models, spatial topic models, etc. Second, this framework easily allows dynamic social networks and/or various temporal effects of users intrinsic

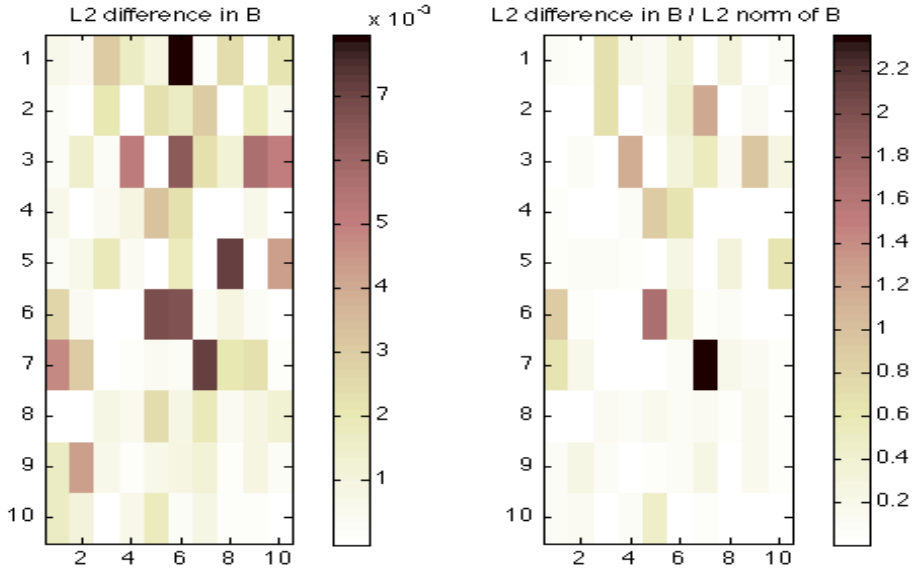

Figure 3.3: Heatmap of L^2 differences (absolute and relative) between entries of true B and estimated B.

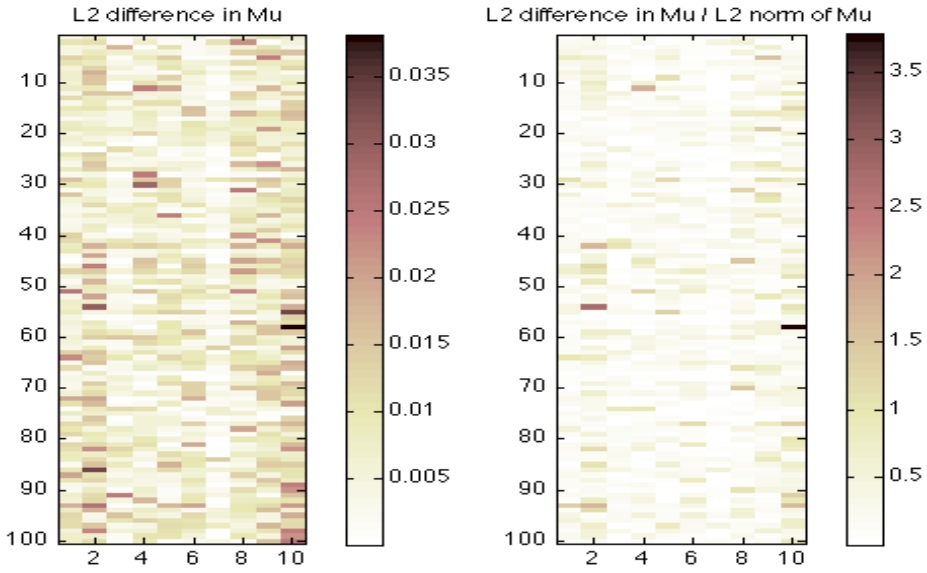

Figure 3.4: Heatmap of L^2 differences (absolute and relative) between entries of true μ and estimated μ.

rate of diffusion to be investigated and discovered.

Our estimation algorithms do not depend on the temporal kernel of the underlying Hawkes process, and a variety of kernels can be used to model different kinds of temporal interactions, for example: close-range interactions with the exponential kernel and long-range interactions with power law kernels. The estimation algorithms remain robust and fast no matter what, and parametric

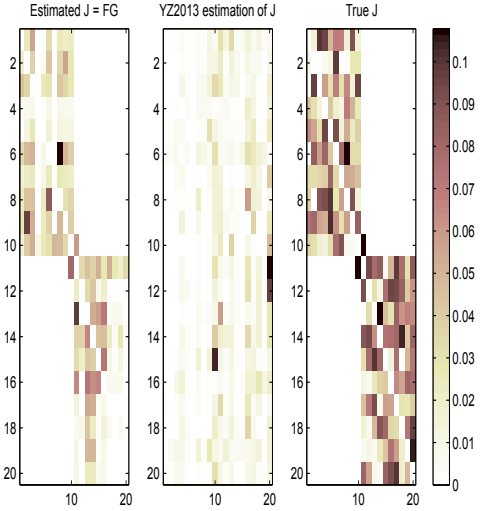

Figure 3.5: Left: our proposed estimation. Center: estimation following [310]. Right: true J.

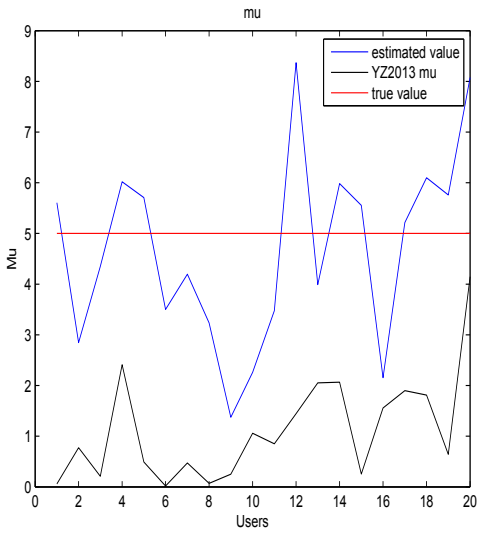

Figure 3.6: Plot of μ for comparison with [310].

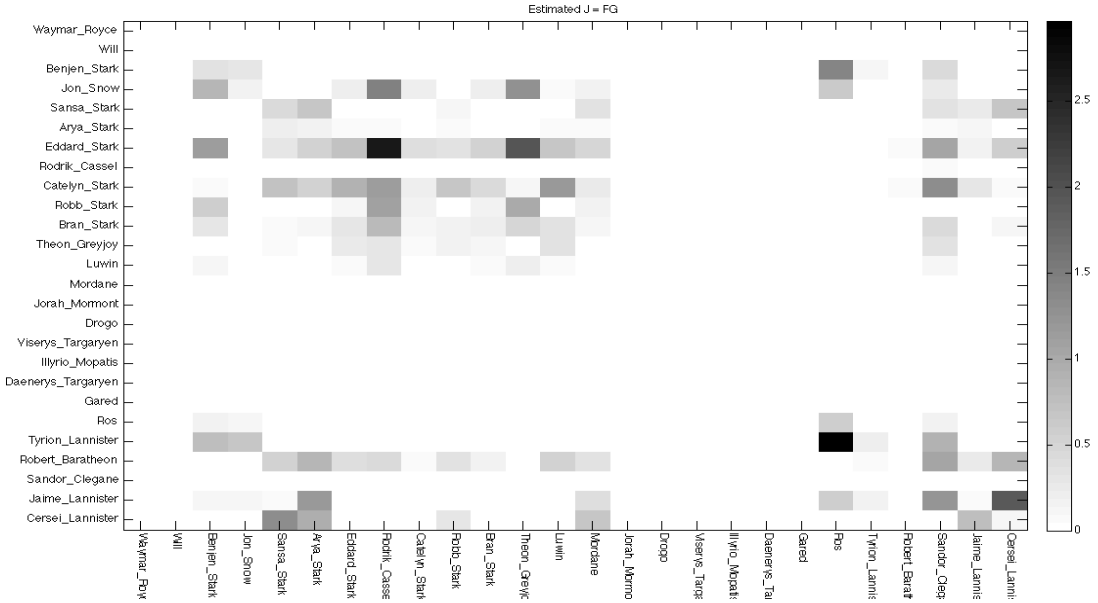

Figure 3.7: Game of Thrones influence heatmap.

estimation algorithms (or even nonparametric ones) for these temporal kernels may also be coupled in our estimation procedure.

The multiplicative updates stemming from the nonnegative tensor factorization are also appealing: the multiplicative updates derived from the optimization problem are easy to implement, even in a distributed fashion - they are basically matrix products and entrywise operations - and the complexity of the algorithm is linear in the data, allowing one to perform estimations in real-life social networks, especially if some of the parameters are already known beforehand.

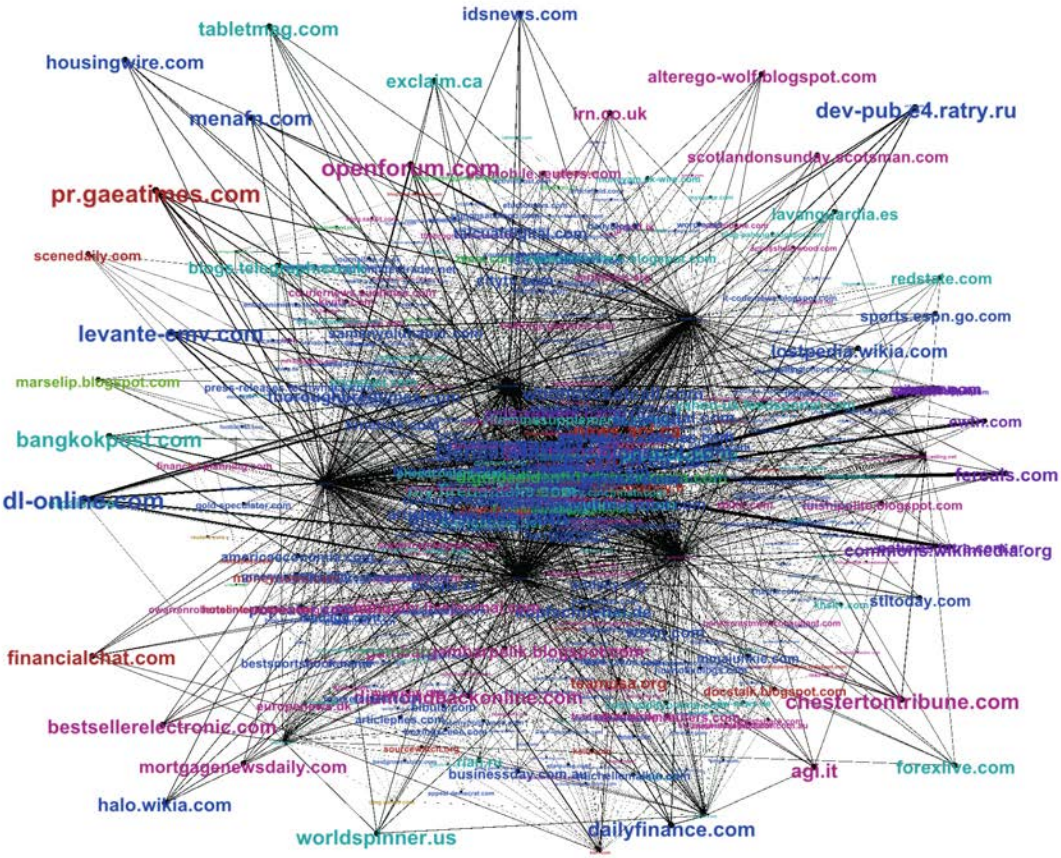

Figure 3.8: Sites influence graph.

One can also notice that by performing a dimensionality reduction during our nonnegative tensor factorization estimation, we not only estimated the influence that users have on one another but we also acquired information on the communities of the underlying social networks, since we were able to factorize the hidden influence graphs. Here, we used heavily the self-exciting model to retrieve the hidden influence graphs, which is different from other graphs generated by different methods; for example, one could weight the communication graph with the number of messages from one user to its neighbors, but by doing so one looses the temporal character. Moreover, the graphs found by performing this kind of technique are under the assumption that messages influence directly other users, which may not be the case. In our Hawkes framework, the influence is a byproduct of the interaction of users and information, and therefore their influence is probabilistic - it may or may not occur at each broadcast.

CHAPTER 4

Trend detection using Hawkes processes

"I don't set trends. I just find out what they are and exploit them."

— Dick Clark

4.1 Introduction

We focus now, in the last chapter of this thesis, on a particular instance of information diffusion: the discovery of trendy topics being disseminated in a social network. The trend detection algorithm developed here stems from a particular instance of the general information diffusion Hawkes framework presented in chapter 3.

Since we are dealing with social networks, we cannot use classical trend detection algorithms [173, 300], as they do not grasp the full relationship between users and contents in the social network. This idea of leveraging social and textual contents is quite recent, with works as [50, 276] shedding some light into the matter.

In order to fully exploit the social ties between users and information in social networks, we base our trend detection algorithm on information diffusion models [120, 169], and more specifically on a Hawkes-based model for information diffusion in social networks [71, 201, 310]. The Hawkes-based model allows: 1) leveraging on the knowledge of the influences between users and contents, 2) to fully explore the real time of broadcasts, 3) leveraging on the knowledge of users intrinsic (or exogenous) rates. Moreover, the Hawkes intensity represents the propensity of users to broadcasts topics at each time, thus serving as proxy for the activity level of topics and users in the social network [99].

We assume that there exist different topics being disseminated in a social network and we employ the Hawkes process to count the number of broadcasts of these topics by each user in the social network. We say that a topic is *trendy* if it has a rapid increase in its broadcasting Hawkes intensity. These topic intensities are combinations of the users broadcasting intensities, where each user contributes to the topic intensities with a measure of his impact on the network, proportional to his network outgoing eigenvector centrality [234]. A trendy topic has then a burst in its broadcasting in the network, which corresponds to an increase in its broadcasting intensity, or a *peak*. Our algorithm thus seeks the "peaks" in the intensity of the underlying Hawkes process in order to determine the topics that are most likely to be trendy.

The difference between the proposed trend detection algorithm and one that looks solely at the topics with the largest number of broadcasts is that we aim to detect those topics that are trendy but

do not necessarily have a large number of broadcasts. Indeed, the most straightforward approach would be to look at the point process intensities and choose those topics with the highest intensities. The approach used in this chapter is different: we do not compare topics between themselves, but rather compare the topic intensities against their expected maximum values at each time, meaning that topics that do not have yet large intensities can indeed be trendy. Still, our algorithm is also able to capture the trendiness coming from large intensities.

The proposed method bares some resemblance with classical works on trend detection. For example, as mentioned above, our algorithm uses Hawkes processes [140, 208] to model the broadcasting/posting times of messages in a social network, which is similar to the infinite-state automaton approach of Kleinberg [173]; the difference between both approaches is how to deal with the intensity stemming from the broadcasting activity: while Kleinberg searches the periods in time with a high frequency of broadcasts about similar contents, we study a Hawkes intensity for broadcasts about contents that can increase even by broadcasted messages about different ones. Since the influences of users and topics in our information diffusion Hawkes model generates correlation in broadcasts between different contents, the work in this chapter also relates through the underlying Hawkes intensity to the work of Wang *et al.* [300], where the authors propose a probabilistic algorithm that discovers correlated bursty patterns and their periods across text streams; the main difference besides the underlying information diffusion model is that we assume the broadcasts to be about specific predefined topics, whereas Wang *et al.* use text mining techniques to unravel the topics, defined as probabilities over vocabularies.

In comparison to other works on trend detection in social networks, our framework resembles the one proposed by Cataldi *et al.* [50], where the authors devise an algorithm to detect real-time emerging topics in Twitter firstly by extracting the contents of the tweets with a model for their life cycle and secondly by considering the social importance of the sources of the tweets, using the Page Rank algorithm. It also resembles the one proposed by Takahashi *et al.* [276], where the authors derive an algorithm focusing on the social aspects of social networks by dynamically generated links between users, and propose a stochastic model for behavior of a generic social network user, detecting the emergence of a new topic. Again, the major difference between these works and this chapter is the underlying model of broadcasts and the fact that our methodology does not rely on text mining techniques since the content of each broadcasted message is assumed to be already labeled.

To the best of our knowledge, our proposal is the first trend detection algorithm that uses point processes and stochastic control techniques. These techniques are successfully used in many other fields, and are complementary tools to machine learning and text mining techniques, hence providing more diversified treatments for this kind of problem.

The remainder of this chapter is organized as follows. In section 4.2, we recall the adopted model of information diffusion in the social network using Hawkes processes. In section 4.3, we define trendiness in our context, detail our trend detection algorithm and derive the trend indices for topics of messages broadcasted in the social network. In section 4.4, we illustrate our algorithm using two different datasets. Section 4.5 eventually concludes the chapter.

4.2 Information diffusion

We start the theoretical study of our trend detection algorithm by adopting a model for information diffusion in social networks. This model is based on an instance of the Hawkes framework for information diffusion studied in chapter 3.

4.2.1 The Hawkes-based model

As discussed during this thesis, Hawkes-based information diffusion models are widely adopted to model information diffusion in social networks [71, 201, 310]. This is due to several reasons, which are nonexhaustively listed here:

- They are point processes [74], and as such they are designed to model discrete events in networks such as posting, sharing, tweeting, liking, digging, etc.
- Hawkes processes are self-excited processes, i.e., the probability of a future event increases with the occurrence of past events.
- They possess a simple and linear structure for their intensity (the conditional expectation of an occurrence of an event, at each time).
- They present simple maximum likelihood formulas [74, 241], which facilitates a maximum likelihood estimation of the parameters.
- A linear Hawkes process can be seen as a Poisson cluster process [141], which permits the distinction of two regimes: a stationary (or stable) regime in which the intensity processes has a stationary and nonexplosive version, and a nonstationary (or unstable) regime, in which the process has an unbounded number of events (see [140, 44] for details).
- It easily allows extensions from the basic model, such as multiple social networks [172], dynamic/temporal networks [154], seasonality and/or time-dependence for the intrinsic diffusion rate of users [292], etc.

Thus, after listing the properties of Hawkes processes that are interesting when modeling information diffusion in social networks, we restate a detailed description of the adopted information diffusion model in this paper, which is the information diffusion model of subsection 3.2.1.

Again, we represent our social network as a communication graph $G = (V, E)$, where V is the set of users with cardinality $\sharp V = N$ and E is the edge set, i.e., the set with all the possible communication links between users, as in chapter 3. We assume this graph to be directed and weighted, and coded by an inward adjacency matrix J such that $J_{i,j} > 0$ if user j is able to broadcast messages to user i, or $J_{i,j} = 0$ otherwise. If one thinks about Twitter, $J_{i,j} > 0$ means that user i follows user j and receives the news published by user j in his or her timeline.

We assume that users in this social network broadcast messages (post, share, comment, tweet, retweet, etc.) during a time interval $[0, \tau]$. These messages represent information about K predefined[1] topics (economics, religion, culture, politics, sports, music, etc.), and at each event the broadcasted message concerns one and only one specific topic among these K different ones.

When broadcasting, users may influence others to broadcast. For example: when tweeting, the user's followers may find the tweet interesting and retweet it to their friends and followers, generating then a cascade of tweets.

We assume that these influences are divided into two categories: user-user influences and topic-topic influences. For example, during these retweeting cascade, users may react differently to the content of the tweet in question, which of course may imply a different influence of this particular tweet among users. By the same token, the followers in question may respond differently depending on the broadcaster, since people influence others differently in social networks.

1. In our work, we rely on text mining techniques only to classify the broadcasted messages into different topics.

The influences are coded by the $N \times N$ matrix J and the $K \times K$ matrix B, such that $J_{i,j} \geq 0$ is the (possible) influence of user i over user j and $B_{c,k} \geq 0$ is the (possible) influence of topic c over topic k.

In light of this explanation, we assume that the cumulative number of messages broadcasted by users is a linear Hawkes process X, where $X_t^{i,k}$ represents the cumulative number of messages of topic k broadcasted by user i until time $t \in [0, \tau]$.

Let $\mathcal{F}_t = \sigma(X_s,\ s \leq t)$ be the filtration generated by the Hawkes process X. Our Hawkes process is then a $\mathbb{R}^{N \times K}$ point process with intensity $\lambda_t = \lim_{\delta \searrow 0} \mathbb{E}[X_{t+\delta} - X_t | \mathcal{F}_t]/\delta$ defined as

$$\lambda_t^{i,k} = \mu^{i,k} + \sum_j \sum_c J_{i,j} B_{c,k} \int_0^{t-} \phi(t-s) dX_s^{j,c},$$

where $\mu^{i,k} \geq 0$ is the intrinsic (or exogenous) intensity of the user i for broadcasting messages of topic k and $\phi(t)$ is a nonnegative causal kernel responsible for the temporal impact of the past interactions between users and topics, satisfying $||\phi||_1 = \int_0^\infty \phi(u) du < \infty$.

The intensity can be seen in matrix form as

$$\lambda_t = \mu + J(\phi * dX)_t B, \tag{4.1}$$

where $(\phi * dX)_t$ is the $N \times K$ convolution matrix defined as $(\phi * dX)_t^{i,k} = \int_0^t \phi(t-s) dX_s^{i,k}$.

Remark: This chapter is not concerned with the estimation of the Hawkes parameters μ, J and B, for which we redirect the reader to subsection 3.3.1 of chapter 3.

4.2.2 Stationary regime

As already mentioned in subsection 4.2.1, one of the main properties of linear Hawkes processes is that they have a narrow link with branching processes with immigration [141], which gives us the following result (whose proof is well explained in [140, 44] and in chapter 12 of [74]):

Lemma 24. *We have that the linear Hawkes process X_t admits a version with stationary increments if and only if it satisfies the following stability condition*[2]

$$sp(J) sp(B) ||\phi||_1 < 1. \tag{4.2}$$

4.3 Discovering trendy topics

After defining in detail the adopted information diffusion framework serving as foundation for our trend detection algorithm, we continue towards the real goal of this paper: *to derive a Hawkes-based trend detection algorithm.*

The proposed algorithm takes into consideration the entire history of the Hawkes process X_t for $t \in [0, \tau]$ and makes a prediction for the trendiest topics at time τ, based on trend indices $\mathcal{I}^k,\ k \in \{1, 2, \cdots, K\}$. It consists of the following steps:

1. Perform a temporal rescaling of the intensity following the theory of nearly unstable Hawkes processes [162], which gives a Cox-Ingersoll-Ross (CIR) process [69] as the limiting rescaled process.

2. Where for a squared matrix A we denote by $sp(A)$ its spectral radius, i.e., $sp(A) = \sup\{|\lambda| \mid det(A - \lambda \mathbb{I}) = 0\}$.

2. Search the expected maxima of the rescaled intensities for each topic $k \in \{1, 2, \cdots, K\}$, with the aid of the limit CIR process. This task is achieved by solving stochastic control problems following the theory developed in [97], which measure the deviation of the rescaled intensities with respect to their stationary mean.

3. Generate from each control problem a time-dependent index $\mathcal{I}_t^k$, which measures the peaks of topic k during the whole dissemination period $[0, \tau]$. We create then the trend indices $\mathcal{I}^k = \int_0^\tau \mathcal{I}_t^k dt$ for each topic $k \in \{1, 2, \cdots, K\}$.

4.3.1 Trendy topics and rescaling

As our algorithm is based on the assumption that *a trendy topic is one that has a rapid and significant increase in the number of broadcasts*, a major tool in the development of this trend detection algorithm is the rescaling of nearly unstable Hawkes processes, developed by Jaisson and Rosenbaum in [162].

As already mentioned in section 4.2, Hawkes processes possess two distinct regimes: a *stable* regime, where the intensity λ_t possesses a stationary version and thus the number of broadcasts remains at most linear, and an *unstable* regime where the number of broadcasts increases in a superlinear fashion.

The intuition behind the rescaling is the following: since we want to measure topics that have a burst in the number of broadcasted messages, we place ourselves between the stable and unstable regimes, where the stability equation (4.2) is satisfied but barely, i.e., $sp(J)sp(B)||\phi||_1 = 1 - \frac{\lambda}{\tau}$ for $\lambda > 0$, and where there exists a drastic change in the behavior of the broadcasts - a Hawkes process satisfying this property is called *nearly unstable* [162]. By placing ourselves in the stable regime, the Hawkes process still possesses a limited number of broadcasted messages, but as we approach the unstable regime, the number of broadcasted messages increases (which could represent trendiness). Our trend detection algorithm uses hence this rationale in order to transform the Hawkes intensity λ_t into a Brownian diffusion, for which stochastic control techniques exist and are easy to implement.

The rescaling works thus in the following fashion: as the trendy data has a large number of broadcasts, we artificially "push" the Hawkes process X to the unstable regime when estimating the parameters μ, B, J and ϕ, in order to accommodate this large quantity of broadcasts. Then, we perform a rescaling to the intensity λ_t, which converges in law when $\tau \to \infty$ to a one-dimensional Cox-Ingersoll-Ross (CIR) process (see theorem 2), whose deviation to the stationary mean is studied using stochastic control techniques, or more precisely, by detecting its expected maxima [97].

Remark: As there are several ways to rescale the intensity λ_t and obtain a nontrivial limit behavior, we have chosen to use the framework of [162] because their rescaling transforms λ_t into a mean-reverting Brownian diffusion, for which there exist detailed studies about finding its expected maxima, such as [97].

Remark: In order to find the most appropriate nearly unstable regime for the Hawkes process X, the choice of the time horizon τ is crucial, as it determines the timescale of the predicted trends. It means that if ones uses τ measured in seconds, the prediction considers what happens in the seconds after the prediction period $[0, \tau]$, if one uses τ measured in days, the prediction considers what happens in the next day or days after the prediction period $[0, \tau]$, etc.

4.3.2 Topic trendiness

We recall the definition of *trendiness* in our context of information diffusion: *a trendy topic is one that has a rapid and significant increase in the number of broadcasts.*

Although this idea is fairly simple, care must be taken: the definition must take into consideration the users in question, since users do not affect it in the same way. For example: if Barack Obama tweets about climate change, one may assume that climate change may become a trendy topic, but if an anonymous user tweets about the same topic, one has less argument to believe that the topic will become trendy. By the same token, if a group composed of many people start tweeting about the latest iPhone, one may consider it a trendy topic, but if only a small group of friends starts tweeting about it, again, one may not be inclined to think so.

Let us discuss it in more details: since the intensity λ_t is associated with the expected increase in broadcasts at time t, we use λ_t as base measure for the trendiness. Moreover, by the previous paragraph, we must also weight the intensity λ_t with a user-network measure responsible for the impact of users on the network. In our case, this user-network measure is the outgoing network eigenvector centrality of users [234].

Mathematically speaking, let v^T be the left-eigenvector of the user-user interaction matrix J, related to the leading[3] eigenvalue $\nu > 0$. Since v is the leading eigenvector of J^T - the outward weighted adjacency matrix of the communication graph in our social network - it represents the outgoing centrality of the network (also known as eigenvector centrality, similar to the PageRank algorithm [117]) and consequently the users' impact on the network, as desired.

Multiplying Eqn. (4.1) in the left by v^T we have that

$$\begin{aligned} v^T\lambda_t &= v^T\mu + v^T J(\phi * dX)_t B \\ &= v^T\mu + \nu v^T(\phi * dX)_t B \\ &= v^T\mu + \nu(\phi * v^T dX)_t B. \end{aligned}$$

Define $\tilde{X}_t = X_t^T v$, $\tilde{\lambda}_t = \lambda_t^T v$ and $\tilde{\mu} = \mu^T v$, where they all belong to $\mathbb{R}^K$. Transposing the above equation we have the topics intensity

$$\tilde{\lambda}_t = \tilde{\mu} + \nu B^T(\phi * d\tilde{X})_t. \tag{4.3}$$

The intensity $\tilde{\lambda}_t$ of the stochastic process $\tilde{X}_t$ has its k^{th} coordinate given by

$$\tilde{\lambda}_t^k = \sum_{i=1}^{N} \lambda_t^{i,k} v_i, \tag{4.4}$$

which means that it represents a topic as a weighted sum by users, where the weights are given by each user impact on the social network.

By reference to the previous Obama example: since Obama has assumedly a large v coefficient (he has a large impact on the network), a topic broadcasted by him should be more inclined to be trendy, and thus have a potentially large increase in $\tilde{X}_t$; on the other hand, if a topic is broadcasted by some unknown person, with a small coefficient v, it will almost not affect the topic intensity $\tilde{\lambda}_t$.

Since $\tilde{X}_t$ is a linear combination of point processes, the increase at time t in $\tilde{X}_t$ can be measured by its intensity $\tilde{\lambda}_t$. *Consequently, we adopt $\tilde{\lambda}_t^k$ as surrogate for topic k trendiness at time t.*

3. This left-eigenvector v^T has all its entries nonnegative, together with the eigenvalue $\nu \geq 0$, by the Perron-Frobenius theorem for matrices with nonnegative entries, without the need of further assumptions. However, we assume without loss of generality that $\nu > 0$, which can be easily avoided during the estimation.

4.3.3 Searching the topic peaks by rescaling

Our algorithm is concerned with the detection of trendy topics at the final diffusion time τ, taking into consideration all the diffusion history in $[0, \tau]$. This means that our goal is to find topics that will possibly have more broadcasts after time τ than they should have, if one looks at their broadcast history in $[0, \tau]$. With that in mind, we say that topic k has a *peak* at time t if its topic intensity $\tilde{\lambda}_t^k$ achieves its maximum expected intensity at time t, which will be determined by Eqn. (4.9).

Since the influences $\nu(\phi * d\tilde{X})_t B$ are always nonnegative in Eqn. (4.3), we can only find peaks when $\tilde{\lambda}_t^k$ is greater than or equal to its intrinsic mean $\tilde{\mu}^k$. Moreover, one can notice that our definition does not take directly into consideration comparisons between topics, i.e., our definitions of trendiness and of peaks are *relative*, although there exist interactions between topics through the topic-topic influence matrix B.

We continue to the formal derivation of the rescaling, which is performed under the following technical assumption[4]:

Assumption 2. *The topic interaction matrix B can be diagonalized into $B = PDP^{-1}$ (where P is the matrix with the eigenvectors of B and D is a diagonal matrix with the eigenvalues of B) and B has only one maximal eigenvalue.*

Moreover, we assume without loss of generality that $D_{i,i} \geq D_{i+1,i+1}$ and that the largest eigenvalue is $D_{1,1} > 0$ (again, by the Perron-Frobenius theorem, since B has nonnegative entries).

Let us use, for simplicity, exponential kernels, i.e., $\phi(t) = e^{-\omega t}\mathbb{I}_{\{t>0\}}$, where $\omega > 0$ is a parameter that reflects the heaviness of the temporal tail. This means that a larger ω implies a lighter tail, and a smaller temporal interaction between broadcasts.

This choice of kernel function implies that our rescaling uses only one degree of freedom - the timescale parameter ω. It is then quite understandable that with just one degree of freedom we can only have one nontrivial limit behavior for our rescaled topic intensities $\frac{\tilde{\lambda}_{\tau t}^k}{\tau}$. This behavior is thus dictated by the leading eigenvector of B when rescaling. This argument further supports assumption 2.

4.3.3.1 Rescaling the topic intensities

Using the decomposition $B = PDP^{-1}$, where D is a diagonal matrix with the eigenvalues of B, we have that Eqn. (4.3) can be written as

$$\tilde{\lambda}_t = \tilde{\mu} + \nu(P^{-1})^T D^T P^T (\phi * d\tilde{X})_t,$$

which when multiplied by P^T by the left becomes

$$\begin{aligned} P^T\tilde{\lambda}_t &= P^T\tilde{\mu} + \nu D^T P^T(\phi * d\tilde{X})_t \\ &= P^T\tilde{\mu} + \nu D(\phi * d(P^T\tilde{X}))_t. \end{aligned}$$

4. The assumption that B can be diagonalized is in fact a simplifying one. One could use the Jordan blocks of B, on the condition that there exists only one maximal eigenvalue. This assumption is verified if, for example, the graph associated with B is strongly connected; which means that every topic influences the other topics, even if it is in an indirect fashion (by influencing topics that will, in their turn, influence other topics, and so on). One can also develop a theory in the case of multiple maximal eigenvalues for B, but it would be much more complicated as the associated stochastic control problem (as in [97]) has not yet been solved analytically, hence numerical methods should be used.

Defining $\chi_t = P^T \tilde{X}_t$, $\varphi_t = P^T \tilde{\lambda}_t$ and $\vartheta = P^T \tilde{\mu}$, we have that χ_t is a K-dimensional stochastic process with intensity

$$\varphi_t = \vartheta + \nu D(\phi * d\chi)_t.$$

Under assumption 2, we have

$$\varphi_t^k = \vartheta^k + \nu D_{k,k}(\phi * d\chi^k)_t, \tag{4.5}$$

where φ_t^k are uncoupled one-dimensional stochastic processes.

Now, following [162], we rescale φ_t by "pushing" the timescale parameter ω to the unstable regime of $\tilde{X}_t$, so as to obtain a nontrivial behavior (peak) for the intensity $\tilde{\lambda}_t$, if any. In light of lemma 24 and assuming an exponential kernel $\phi(t) = e^{-\omega t}.\mathbb{I}_{\{t>0\}}$, we have that the timescale parameter ω satisfies, for some $\lambda > 0$, $\tau(1 - \frac{\nu D_{1,1}}{\omega}) \sim \lambda$ when $\tau \to \infty$, which implies (we assume without loss of generality that $\tau > \lambda$)

$$\omega \sim \frac{\tau \nu D_{1,1}}{(\tau - \lambda)}. \tag{4.6}$$

The rescaling stems from the next theorem (the one-dimensional case is proven in theorem 2.2 of [163]), which is proven in appendix D, subsection D.1.4:

Theorem 2. *Let assumption 2 be true, the temporal kernel be defined as $\phi(t) = e^{-\omega t}.\mathbb{I}_{\{t>0\}}$, let $\rho = ((P^{-1})_{1,1}, \cdots, (P^{-1})_{1,K})$ be the leading left-eigenvector of B, $\tilde{v}$ be the leading right-eigenvector of J, and define $\pi = (\sum_k (P_{k,1})^2 \rho_k)(\sum_i v_i^2 \tilde{v}_i)$.*

If $\omega \sim \frac{\tau \nu D_{1,1}}{(\tau - \lambda)}$ when $\tau \to \infty$, then the rescaled process $\frac{1}{\tau}\varphi_{\tau t}^1$ converges in law, for the Skorohod[5] topology in $[0,1]$, to a CIR process C^1 satisfying the following stochastic differential equation (SDE)

$$\begin{cases} dC_t^1 = \lambda \nu D_{1,1}(\frac{\vartheta^1}{\lambda} - C_t^1)dt + \nu D_{1,1}\sqrt{\pi}\sqrt{C_t^1}dW_t \\ C_0^1 = 0, \end{cases} \tag{4.7}$$

where W_t is a standard Brownian motion.

Moreover for $k > 1$, the rescaled processes $\frac{1}{\tau}\varphi_{\tau t}^k$ converge in law to 0, for the Skorohod topology to in $[0,1]$.

As a result, we are only interested in the CIR process C^1, since it is the only one that possesses a nontrivial behavior. One can clearly see that, since a CIR process is a mean-reverting one, C^1 mean-reverts to the stationary expectation $\overline{\mu} = \frac{\vartheta^1}{\lambda}$. As already discussed in subsection 4.3.3, if one wants to capture some trend behavior one must see this process above its stationary expectation $\overline{\mu}$, i.e., one must study the process $C_t = C_t^1 - \overline{\mu}$.

By Eqn. (4.7), one easily has that $C_t = C_t^1 - \overline{\mu}$ satisfies the following SDE:

$$dC_t = -\lambda \nu D_{1,1} C_t dt + \nu D_{1,1}\sqrt{\pi}\sqrt{C_t + \overline{\mu}}dW_t. \tag{4.8}$$

Remark: A way of pushing the Hawkes process to the instability regime, when estimating the matrices μ, J and B, is to put the timescale parameter ω near the stability boundary given by Eqn. (4.2).

5. The Skorohod topology in a given space is the natural topology to study càdlàg processes, i.e., stochastic processes that are right-continuous with finite left limits. This topology has the goal to define convergence on cumulative distribution functions and stochastic processes with jumps. See [27] for a formal definition.

4.3.3.2 The trend index

After rescaling the $\varphi_t = P^T\tilde{\lambda}_t$, we effectively search for the peaks in $\tilde{\lambda}_t$ using the framework developed by Espinosa and Touzi [97] dedicated to search for the maximum of scalar mean-reverting Brownian diffusions.

For that goal, we define trend indices $\mathcal{I}_t^k$ as the measure, at each time instant $t \in [0, \tau]$, of how far is the intensity $\tilde{\lambda}_t^k$ from its peak, where a peak is represented by a maximum of $\tilde{\lambda}_t^k$ in the sense of [97]. To do so, we use the fact that $\tilde{\lambda}_t = (P^{-1})^T\varphi_t$ to determine the limit behavior of $\frac{\tilde{\lambda}_{\tau t}^k}{\tau}$, namely $\tilde{\lambda}_t^{k,\infty}$, as

$$\tilde{\lambda}_t^{k,\infty} = \sum_j P_{j,k}^{-1} C_t^j = P_{1,k}^{-1} C_t^1 = P_{1,k}^{-1}(C_t + \overline{\mu}),$$

where P is the eigenvector matrix of B in assumption 2 and C_t^k are the rescaled CIR processes in theorem 2.

Hence, in order to find our intensity peaks, we consider for each topic k the following optimal stopping problem

$$V_k = \inf_{\theta \in \mathcal{T}_0} \mathbb{E}[\frac{(P_{1,k}^{-1})^2}{2}(C_{T_0}^* - C_\theta)^2], \tag{4.9}$$

where $C_t^* = \sup_{s \leq t} C_s$ is the running maximum of C_t, $T_y = \inf\{t > 0 \mid C_t = y\}$ is the first hitting time of barrier $y \geq 0$ and $\mathcal{T}_0$ is the set of all stopping times θ (with respect to C) such that $\theta \leq T_0$ almost surely, i.e., all stopping times until the process C reaches 0.

By the theory developed in [97], one has optimal barriers γ^k relative to each problem V_k. *A barrier represents the peaks of the intensities, i.e., if the CIR process C touches the optimal barrier γ^k, it means that we have found a peak for topic k.*

The authors show that the free barriers γ^k have two monotone parts; first a decreasing part $\gamma_\downarrow^k(x)$ and then an increasing part $\gamma_\uparrow^k(x)$, which are found by solving the ordinary differential equations (ODE) (5.1) and (5.15) in [97], respectively[6].

We are now able to define for each time $t \leq T_0$, the temporal trend indices $\mathcal{I}_t^k$ as

$$\mathcal{I}_t^k = \begin{cases} \psi^+(\tau - t, C_t - \gamma^k(C_t)) & \text{if } t < \tau \text{ and } C_t \geq 0, \\ \psi^-(\tau - t, C_t - \gamma^k(C_t)) & \text{if } t < \tau \text{ and } C_t < 0, \\ \Psi^+(C_\tau - \gamma^k(C_\tau)) & \text{if } t = \tau \text{ and } C_t \geq 0, \\ \Psi^-(C_\tau - \gamma^k(C_\tau)) & \text{if } t = \tau \text{ and } C_t < 0, \end{cases}$$

where $\psi^{+/-}$ are decreasing in time (the first variable), increasing in space (the second variable) functions and $\Psi^{+/-}$ are increasing in space functions. We impose $\psi^{+/-}$ as decreasing functions of time because our trend detection algorithm is to determine the trendy topics at time τ, the end of the estimation time period. Thus the further we are in the past (measured by $\tau - t$), the less influence it must have in our decision, and consequently in our trend index. By the same token, $\psi^{+/-}$ and $\Psi^{+/-}$ must be increasing functions in space because we want to distinguish topics that have higher intensities, and penalize those that have a lower intensity, thus if the intensity is bigger

6. For the CIR case we have by Eqn. (4.8) that the functions α, S and S' defined in [97] are

- $\alpha(x) = \frac{2\lambda x}{\nu D_{1,1}\pi(x+\overline{\mu})}$, $S'(x) = e^{\frac{2\lambda x}{\nu D_{1,1}\pi}}(\frac{x}{\overline{\mu}} + 1)^{-\frac{2\lambda\overline{\mu}}{\nu D_{1,1}\pi}}$ and
- S is a linear combination of a suitable transformation of the confluent hypergeometric functions of first and second kind, M and U, respectively (see [1]), since it must satisfy $S(0) = 0$ and $S'(0) = 1$ (see [186]).

than the optimal barrier, we must give it a bigger index. If, on the other hand, the intensity is smaller than the optimal barrier, even negative in some cases, we must take into account the degree of this separation. One has the liberty to choose the functions ψ and Ψ according to some calibration dataset, which makes the model more versatile and data-driven.

Please note that in the definition of $\mathcal{I}_t^k$, the following factors have been taken into consideration:

- even if the CIR intensity C_t did not reach its expected maximum given by $\gamma^k(C_t)$, we must account for the fact that it may have been close enough,
- reaching the expected maximum is good, but surpassing it is even better. So we must not only define a high trend index if C_t reaches the expected maximum given by $\gamma^k(C_t)$, but we must define a *higher* trend index if C_t surpasses these barriers, and
- it is important to *penalize* all the times $t \in [0, \tau]$ that the intensity C_t becomes negative, i.e., the intensities $\tilde{\lambda}_t^k$ become smaller than their stationary expectation.

The trend indices $\mathcal{I}^k$ are thus defined as

$$\mathcal{I}^k = \int_0^\tau \mathcal{I}_t^k dt.$$

Remark: One could be also interested in not only tracking the relative trendiness of each topic with respect to their maxima, but also the *absolute* trendiness of topics with respect to each other. In this case, one may define the trend indices $\tilde{\mathcal{I}}_t^k$ as

$$\tilde{\mathcal{I}}_t^k = \mathcal{I}_t^k + a(\tau - t)\tilde{\lambda}_t^{k,\infty} = \mathcal{I}_t^k + a(\tau - t)P_{1,k}^{-1}(C_t + \overline{\mu}),$$

where $a(\tau - t) \geq 0$ are nonincreasing functions of time (again, in order to give a bigger influence to the present compared to the past). The absolute trendiness of topics can be explained as follows: Lady Gaga may be not trendy according to our definition, if for example people do not tweet *as much as expected* about her at the moment, but she will probably still be trendier than a rising-but-still-obscure Punk-Rock band. In this case, the relative trend index $\mathcal{I}^k$ of Lady Gaga is not that big as compared to the relative trend index of the Punk-Rock band. However, the absolute trend index $\tilde{\mathcal{I}}^k$ of Lady Gaga will surely be bigger than the absolute trend index of the Punk-Rock band, if the function $a(\tau - t)$ is large enough. The function $a(\tau - t)$ controls which behavior one wants to detect, the relative or the absolute trendiness.

Remark: This algorithm is fast, despite the use of numerical discretization schemes for the ODEs. By using the eigenvector centrality of the underlying social network as tool to create our trend indices, we not only use the topological properties of the social network in question but we reduce considerably the dimension of the problem: we only have a one-dimensional CIR process to study. Moreover, the complexity of the algorithm breaks down to three parts: 1) the resolution of the K optimal barrier ODEs, which is of order $\mathcal{O}(\frac{K}{\delta})$ where δ is the time-discretization step, 2) the calculation of the left and right leading eigenvectors of J and B, which can be achieved fairly fast with iterative methods such as the power method, and 3) the matrix product in the calculation of $\overline{\mu}$, which has complexity $\mathcal{O}(NK)$.

4.4 Numerical Examples

We provide in this section two examples where we apply our trend detection algorithm.

Algorithm 3 - Trend detection algorithm

Input: Hawkes process X_t, $t \in [0, \tau]$, matrices J, B and μ
Compute the leading left-eigenvector v^T and eigenvalue ν of J, and the topic intensities $\tilde{\lambda}_t$ following Eqn. (4.3)
Compute the leading right-eigenvector $(P_{11}, \cdots, P_{K1})$, left-eigenvector $(P_{11}^{-1}, \cdots, P_{1K}^{-1})$ and eigenvalue $D_{1,1}$ of B, and the leading right-eigenvector $\tilde{v}$ of J
"Push" $\tilde{\lambda}_t$ to the instability regime following Eqn. (4.6) and calculate the CIR intensity C_t following Eqn. (4.8)
Discretize $[0, \tau]$ into T bins of size $\delta \ll 1$
for $k = 1$ to K **do**
    Get the optimal barrier γ^k in $\{0, \delta, 2\delta, \cdots, (T-1)\delta\}$, following [97]
    for $t = 1$ to T **do**
        Calculate the trend index $\mathcal{I}^k_{(t-1)\delta}$ using the optimal barrier γ^k of the optimal stopping problem (4.9)
    end for
    Calculate the topic trend index $\mathcal{I}^k = \int_0^\tau \mathcal{I}^k_t dt = \delta \sum_{t=1}^T \mathcal{I}^k_{(t-1)\delta}$
end for
Output: Trend indices $\mathcal{I}^k$

The first example is performed on a synthetic near unstable Hawkes processes in a social network using Ogata's thinning algorithm[7] [241] in a time horizon $\tau = 50$. We use 10 topics for the simulation, the last 5 topics not possessing *any topic influence*, i.e., $B_{c,k} = 0$ for all c and $k \in \{6, 7, 8, 9, 10\}$, corresponding to figures 4.1 and 4.2.

The second example is applied to a MemeTracker dataset containing different memes (short distinct phrases) for the 5,000 most active sites from 4 million sites from March 2011 to February 2012[8]. We use the 10 most broadcasted memes, which are:

1. dancing with the stars,
2. two and a half men,
3. sex and the city,
4. rolling in the deep,
5. too big to fail,
6. don't ask, don't tell,
7. i have a dream,
8. i will always love you,
9. the girl with the dragon tattoo,
10. the tree of life.

7. The thinning algorithm simulates a standard Poisson process P_t with intensity $M > \sum_{i,k} \lambda_t^{i,k}$ for all $t \in [0, \tau]$ and selects from each jump of P_t the Hawkes jumps of $X_t^{i,k}$ with probability $\frac{\lambda_t^{i,k}}{M}$, or no jump at all with probability $\frac{M - \sum_{i,k} \lambda_t^{i,k}}{M}$.
8. Data available at http://snap.stanford.edu/infopath.

Table 4.1: Comparison of indices for synthetic dataset.

Topic	1	2	3	4	5	6	7	8	9	10
$\tilde{\mathcal{I}}$	**0.104**	0.088	0.097	0.096	0.0889	0.033	0.033	0.033	0.033	0.0334
Nb of Posts	47640	41770	**56368**	51039	56097	55252	48105	48096	43882	53580

Table 4.2: Comparison of indices for MemeTracker dataset.

Meme	1	2	3	4	5	6	7	8	9	10
$\tilde{\mathcal{I}}$	0.002	0.001	0.004	0.009	0.005	0.003	0.007	0.009	**0.0159**	0.011
Nb of Posts	1768	**1925**	1406	1537	1578	1871	1746	1562	1344	1499

In this numerical example, each meme plays the role of a topic in our theoretical model of section 4.2. We use the maximum likelihood estimation procedure for the parameters, as detailed in the subsection 3.2.1 of chapter 3.

For both examples, we shall illustrate how our method is able to detect the trendiness of each topic or meme according to the index $\mathcal{I}$, and that the highest trendiness does not necessarily correspond to the topic or meme which has the highest number of broadcasts.

For both examples, figures 4.1 and 4.3 plot the scaled topic intensities $\frac{\tilde{\lambda}_{\tau t}}{\tau}$ as a function of time, and figures 4.2 and 4.4 plot the cumulative number of broadcasts about each topic as a function of time, i.e., $\overline{X}_t^k = \sum_i X_t^{i,k}$. Furthermore, we compute in tables 4.1 and 4.2 the trend indices $\tilde{\mathcal{I}}^k$ and the total number of broadcasts for both examples. We use for the trend indices calculation the following functions $\psi^{+/-}(t,x) = \frac{e^{2x}}{t+1}$, $\Psi^{+/-}(x) = 2x$ and $a(t) = \frac{1}{t+1}$, as explained in subsubsection 4.3.3.2.

In reference to table 4.1, one can see that the trend index for topic 1 is the highest, even though it does not possess the highest number of broadcasts, which is held by topic 3. The reason is that in the synthetic dataset, topic 1 has the largest topic intensity.

In reference to table 4.2, one can see that meme 9 shows higher trendiness than the other memes, even though it does not possess the highest intensity and it possesses the smallest total number of broadcasts; it is then followed by memes 10 and 8 in second and third place, respectively. The reason is similar for all of them: they possess larger and more frequent "peaks" of intensity compared to other memes, occurring at times closer to the prediction instant τ, as depicted in figure 4.3.

A different phenomenon occurs for meme 2, which has the highest total number of broadcasts but the least trendiness. Since most of the broadcasts of meme 2 occur very early in time, the peak of intensity related to this increase in the number of broadcasts has little impact in the trend index, which takes more into account broadcasts that happen near the prediction instant τ. Thus, as meme 2 does not have significant peaks in intensity near the prediction instant τ, it receives a lower trend index. *Both phenomena illustrate the difference between our algorithm and one that looks solely to the largest topic intensities and the total number of broadcasts.*

4.5 Conclusion

We have developed in this chapter a trend detection algorithm, designed to find trendy topics being disseminated in a social network. We have assumed that broadcasts of messages in the social network can be modeled by a self-exciting point process, namely a Hawkes process, which takes into

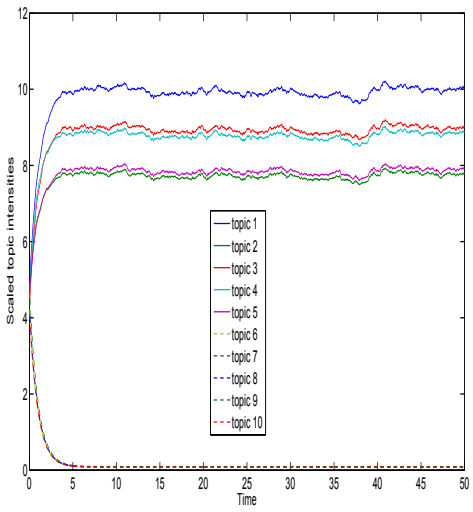

Figure 4.1: Topic intensities $\tilde{\lambda}_t^k = \sum_i \lambda_t^{i,k} v_i$ for the synthetic dataset.

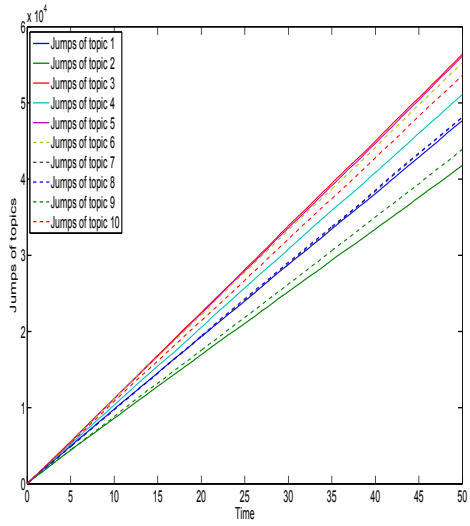

Figure 4.2: Cumulative sum of jumps of topics $\overline{X}_t^k = \sum_i X_t^{i,k}$ for the synthetic dataset.

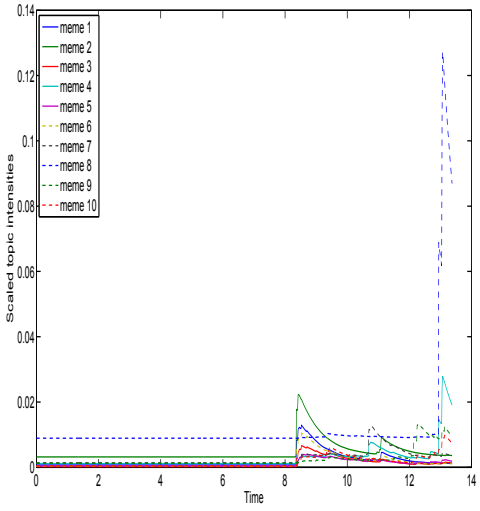

Figure 4.3: Topic intensities $\tilde{\lambda}_t^k = \sum_i \lambda_t^{i,k} v_i$ for the meme tracker dataset.

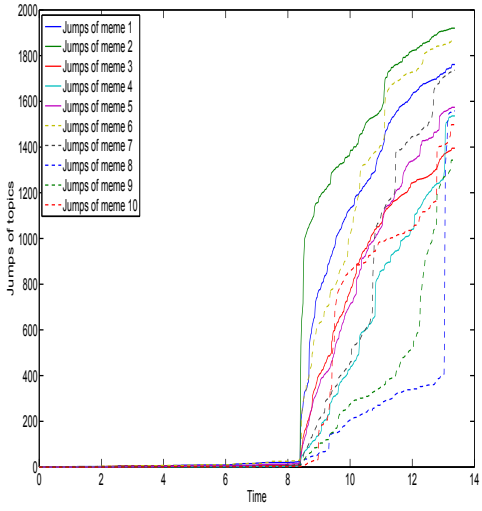

Figure 4.4: Cumulative sum of jumps of topics $\overline{X}_t^k = \sum_i X_t^{i,k}$ for meme tracker dataset.

consideration the real broadcasting times of messages and the interaction between users and topics.

We defined our idea of trendiness and derived trend indices for each topic being disseminated. These indices take into consideration the time between the actual trend detection and the message broadcasts, the distance between the intensity of broadcasting and the maximum expected intensity of broadcasting, and the social network topology. This result is, to the best of our knowledge, the first definition of relative trendiness, i.e., a topic may not be very trendy in absolute number of broadcasts when compared to other topics, but has still rapid and significant number of broadcasts as compared to its expected behavior. Still, one can easily create an absolute trend index for each

topic in our trend detection algorithm, where all one needs to do is use the broadcasting intensities of each topic as surrogates for their trendiness. It is worthy mentioning that these broadcast intensities also take into consideration the social network topology, or more precisely, the outgoing eigenvector centrality of each user, i.e., their respective influences on the social network.

The proposed trend detection algorithm is simple and uses stochastic control techniques in order to derive a free barrier for the maximum expected broadcast intensity of each topic. This method is fast and aggregates all the information of the point process into a simple one-dimensional diffusion, thus reducing its complexity and the quantity of data necessary to the detection - indispensable features if one is concerned with the detection of trends in real-life social networks.

Conclusion

"I am turned into a sort of machine for observing facts and grinding out conclusions."

— Charles Darwin

We have started this thesis with a simple opinion dynamics model. The reason for that is quite concrete: opinion dynamics models were the predecessors of social network analysis and laid the ground for all the nonstandard and complex models that were used and discussed in this thesis.

The goal was to produce an opinion dynamics model that could handle multiple contents and stochastic interactions between agents in a social network. We have thus developed a model where agents could transmit information to their neighbors, the information itself being stochastic - it would depend on the appreciation of the broadcaster towards the contents.

The tools used to analyze this opinion dynamics algorithm were stochastic approximation algorithms (which are used in classical and distributed optimization problems, control theory, machine learning problems, etc.) and game theory, or more precisely stochastic fictitious play (game-theoretical learning paradigms in which agents learn about the possible Nash equilibria by playing consecutive games with the time average of the agents' ancient responses).

We used different choice mechanisms for the content being broadcasted and proved mathematically that when the communication network is undirected this novel opinion dynamics algorithm converges, as time goes by. The limit behavior of this algorithm depends, of course, on the way agents interact - which is summarized by a rationality parameter: when this parameter is very small, agents are more dubious about their preferences and thus the system converges to a state of complete uncertainty - uniform preferences of contents. When this parameter is large, agents become more and more certain about which content to broadcast, and thus the system converges to a clustering of the social network in question, where clusters are composed of agents that broadcast the content they appreciate the most; it allows each cluster to be associated with a specific content.

Using this opinion dynamics algorithm under the assumption that agents are quite certain about which content to broadcast at each time (a large rationality parameter), we created a community detection algorithm that is easily implemented, even in a distributed fashion, and generates communities that can be mathematically analyzed - a fact that is quite rare in the community detection community. The fact that communities can be analyzed opens new frontiers in terms of a mathematically sound definition for communities in social networks, as most of the community detection algorithms are based on heuristics and the discovered communities cannot be analyzed mathematically (empirical validations on the goodness-of-fit of these algorithms must be taken into consideration).

Our algorithm, as already mentioned, is based on a well-studied opinion dynamics algorithm and discovers communities based on a sound mathematical formulation, with the advantage of

being versatile: it can also be used in weighted graphs; the communities found still satisfy the same assumptions, with the additional benefit of incorporating community overlaps without overhead.

This community detection algorithm has two variants, depending on the previous knowledge one has about the social network: by limiting the number of contents in the underlying opinion dynamics algorithm, one may drastically decrease the complexity of the algorithm; this means limiting the maximum number of communities and giving it a more parametric flavor. On the other hand, if one uses the full-fledged algorithm, it becomes completely nonparametric. It means that our community detection algorithm is multiscale, i.e., it can adapt to the desired final resolution of the social network in question.

After better comprehending the interactions of users in social networks due to the creation of our opinion dynamics model and the associated community detection algorithm, we introduced a complete framework for the study of hidden influences on social networks.

The proposed framework is based on self-exciting point processes, the so-called Hawkes processes. This type of stochastic process was developed by Hawkes in the 70's to model earthquakes, which have a positive temporal correlation, i.e., the probability of a future occurrence of an earthquake increases with past earthquake activity. The same effect happens in social networks, or in any conversation for that matter: the probability of posting, reposting, retweeting, sharing or liking increases if some previous action was performed.

With this in mind, we created a Hawkes-based theoretical framework for information diffusion in social networks that models and discovers the hidden influences between users and contents. This framework is general by construction, and adapts to one's needs at ease: it can accommodate user-user interactions, user-topic interactions, topic-topic interactions, multiple social networks or a single social network, static social networks or dynamic/temporal social networks, seasonality or different temporal effects on the users diffusion rate, different temporal interactions (short-range or long-range interactions), predefined topics (known beforehand) or unknown topics (combined use of several topic models for the estimation procedure). As one can see, our framework is sufficiently general to model and estimate most of interesting phenomena occurring on social networks, with simplifications being made only for the sake of reducing the complexity since the theory deals with real-life social networks.

The estimated procedure developed in order to discover these hidden influences and temporal phenomena is based on a maximum likelihood estimation of the underlying Hawkes processes, using nonnegative tensor factorization techniques. Nonnegative tensor factorization techniques are a set of tools developed in the signal and image processing community in order to seek patterns in multidimensional noisy data, for example image and sound decomposition. In our particular case, we adapt the so called multiplicative updates stemming from nonnegative matrix factorization techniques in order to derive a simple, distributed, fast and easy-to-implement estimation framework for the hidden parameters which, due to a dimensionality reduction algorithm, has a linear complexity on the data.

The dimensionality reduction (the factorization of the influence graphs) provides information on the "hidden" communities of the social networks in question, which is a byproduct achieved with no overhead at all. These communities cannot be uncovered by standard social network analysis tools since they represent the "hidden" influence of users over one another, and are different from other measures generated by these tools; for example, one could weight the communication graph with the number of messages from one user to his neighbors, but by doing so one looses the temporal character. Moreover, the graphs found by performing this kind of technique are under the assumption that messages influence directly other users, which may not be the case. In our Hawkes framework, the influence is a consequence of the interaction between users and information, and

therefore it is probabilistic by definition - it may or may not occur at each broadcast.

The last part of this thesis bases itself on one instance of the previously defined Hawkes information diffusion framework, in order to derive a trend detection algorithm for topics in social networks. We have assumed that broadcasts of messages in a social network are modeled by a Hawkes process, which takes into consideration the real broadcasting times of messages and the interaction between users and topics.

We used the influence matrix (which is estimated following the ideas in this thesis) to define the topic diffusion intensities: when someone posts or retweets something, the influence that this person has over the social network must be taken into consideration in order to calculate a meaningful notion of topic trendiness. This influence takes the form of an eigenvector centrality measure, the same sort of measure used in the PageRank algorithm.

We also defined our idea of trendiness: a topic may not be very trendy in absolute number of broadcasts when compared to other topics, but may still have rapid and significant number of broadcasts as compared to its expected behavior. This idea is, to the best of our knowledge, the first definition of relative trendiness. Still, one can easily integrate an absolute trend index for each topic, simply by introducing the broadcasting intensities of each topic as surrogates for their trendiness.

We defined thus trend indices, one for each topic, that aggregate these characteristics in a single value: they take into consideration the time between the actual trend detection and the message broadcasts, the distance between the intensity of broadcasting and the maximum expected intensity of broadcasting, and the social network topology.

The calculation of these indices are a two-step procedure: first, we considered a near-unstable Hawkes framework in order to find a suitable time rescaling for the topic intensities, transforming them into a Cox-Ingersoll-Ross process, which is an Itô diffusion. Second, we adapt an optimal stopping problem previously studied for the detection of the expected maximum of a mean-reverting scalar Itô diffusion, which allows the search for "peaks" in the rescaled topic intensities; the optimal stopping tools used in this detection problem come from the more general stochastic control toolbox, a set of techniques and methods developed to study control problems under uncertainty, and successfully applied in mathematical finance, control theory, continuous-time optimization and decision problems.

The rescaling step is important for two reasons: first, since we look at bursts of broadcasts, the "stretching" of time finds a more appropriate timeframe for the topic intensities themselves. Second, it transforms the intensities in Brownian diffusions, for which there exists a multitude of stochastic control methods.

The adopted optimal stopping theory allows the derivation of free barriers for the maximum expected broadcast intensity of each topic, which are responsible for the calculation of the trend indices. These indices take hence the distance between the topic intensities and their maximum expected intensities, generating (as desired) a relative trend index for each topic. An absolute trend index can be added, by incorporating information about the topic intensities themselves, which gives a more versatile and applicable trend detection algorithm.

This method is fast and aggregates all the information of the point process into a simple one-dimensional diffusion, thus reducing its complexity and the quantity of data necessary to the detection - indispensable features if one is concerned with the detection of trends in real-life social networks.

In times where information is jury, judge and executioner, I cannot imagine more fitting final words than Human.4's Mike A. Lancaster's: "We simply don't have enough data to form a conclusion."

Future work

This thesis tackled the problem of information diffusion in social networks in two ways: we first developed an opinion dynamics model, from which we derived a community detection algorithm, and we second created a Hawkes-based information diffusion framework, which generated a Hawkes-based trend detection algorithm in social networks. Future work on these subjects has several scopes, for which we provide a non-exhaustive list of potential directions.

Regarding the opinion dynamics algorithm of chapter 1, these directions may be

- Our main convergence result, theorem 1, is proven under the assumption that the network is undirected, whereas numerical evidence conjectures that this theorem remains valid for directed networks as well. Thus, a direction of future research may be to prove a more general convergence result for directed networks or construct a counter-example in the negative case.

- A more detailed study of the limit set $\mathcal{F}_\beta^x$ for the case $\beta \gg 1$, as follows: Stochastic fictitious play literature [149, 290] assumes that when $\beta \to \infty$ the limit set $\mathcal{F}_\beta^y$ converges to the finite set $\mathcal{F}_\beta^y = \{y \in (\Delta_K)^N \mid y = f_\infty(D^{-1}Ay)\}$, where

$$\begin{aligned} f_\infty : (\Delta_K)^N &\to (\Delta_K)^N \\ x \mapsto f_\infty^{ik}(x) &= \frac{\mathbb{I}_{\{k \in argmax_l(x^{i,l})\}}}{\sharp argmax_l(x^{i,l})}. \end{aligned}$$

  This is assumed since f_β converges pointwise to f_∞ when $\beta \to \infty$ (a result that seems to be true in the performed numerical simulations). This method is called equilibrium selection [138].

  A partial answer is positively answered by Peter Tiňo in [285, 286], showing in addition that we also have a finite number of fixed points for the softmax function f_β in the vector case (which is equivalent to demanding the matrix A to be the identity matrix in our case).

- Exploit heterogeneity of agents, as in [106]. This can be accomplished by defining for each agent $i \in V$ a specific softmax parameter β_i. We can have thus different groups of agents, one in which all agents have a small softmax parameter and thus are indifferent to contents, another in which agents have a large softmax parameter and thus almost always broadcast the content they appreciate the most, etc.

- Use dynamic [154], random [39], multiplex networks [87, 172], etc. For example, a multiplex network is a network $G = (V, E)$ in which edges have classes: imagine that the union of all social networks is a multiplex with every social network being a class. Nodes may thus have more than one edge between them, which could provide a better model for opinion dynamics with agents belonging to multiple social networks.

About the subsequent community detection algorithm of chapter 2, future directions may be

- Develop theoretical and/or heuristic ways to improve the speed of the algorithm (for example, derive better bounds on the running time T).

- Provide a more rigorous and/or detailed analysis of the impact on the retrieved communities due to the initial condition.

- Create a more systematic way of exploring the multiscale character of the algorithm based on the number of contents K and the softmax parameter β.
- In the Potts model clustering algorithm [29, 230], the softmax parameter β plays the role of "inverse temperature", which is responsible for the granularity of the communities found. The Potts model clustering algorithm possess three distinct phases, for which the temperature responsible for the phase transition can be estimated using mean-field [307] and Markov-Chain Monte-Carlo [298] approaches. One may thus apply these methods to derive better bounds for β.
- Extend the community detection algorithm to dynamic/temporal networks [154], following the opinion dynamics framework of chapter 1.
- Extend the community detection algorithm to multiplex networks [87, 172], following the opinion dynamics framework of chapter 1.

Regarding the second part of this thesis concerning Hawkes-based information diffusion methods, future directions on the trend detection algorithm of chapter 4 may be

- Introduce topic models as in chapter 3.
- Extend the trend detection algorithm to accommodate dynamic networks [154] or multiplex networks [87, 172], following the ideas of chapter 3.
- Introduce seasonality or temporal aspects for the intrinsic rates of diffusion μ, as in chapter 3.
- Introduce different temporal kernel functions ϕ.
- Provide a standard way of "pushing" the Hawkes process to the unstable regime when estimating the Hawkes parameters μ, J, B and ϕ.
- Derive a more data-driven way to determine the time-rescale parameters τ and λ.
- Measure the impact of the temporal kernel ϕ on the trend detection algorithm.
- Measure the impact of the temporal discounting functions $\psi^{+/-}$, the final barrier functions $\Psi^{+/-}$ and the function $a(t)$ on the trend indices.

Appendix

APPENDIX A

Opinion dynamics with $f(P) = P$

In chapter 1 we presented an opinion dynamics model with choice function f the softmax function with parameter β, given by Eqn. (1.1). This choice function stems from the concept of bounded rationality (as seen in chapter 1) and represents the certitude that agents have when choosing a content to broadcast.

This appendix is concerned with a different kind of choice function, the identity function $f(P) = P$. This function represents the broadcast probability of a content $k \in \{1, \cdots, K\}$ to be proportional to the appreciation that agents have for this particular content.

This model differs from the one with a softmax choice function for a simple reason: with the softmax choice function with a parameter $\beta \gg 1$, agents have a much higher probability to broadcast the topics with maximum appreciation, in a superlinear way, due to the exponential function. If $f(P) = P$, agents still have a higher probability to broadcast the contents with maximum appreciation, however this difference is not superlinear, which smooths the evolution of differences in the topics with maximum and average appreciation[1].

As already shown in chapter 2, when $\beta \gg 1$ the opinion dynamics algorithm (1.9) converges to a clustering of the graph $G = (V, E)$, where clusters are composed of agents that broadcast internally the same content. We show in this appendix that when the choice function is $f(P) = P$, the convergence result is quite different: the opinion dynamics algorithm converges to a consensus on each strongly connected component of G.

We first define rigorously a consensus:

Definition 4. *(Reaching Consensus) We say that the stochastic approximation algorithm* (1.9) *reaches a consensus if:*

$$\max_{0 \leq k \leq K} |P_t^{i,k} - P_t^{j,k}| \overset{t \to \infty}{\to} 0, \ \forall i, j \in V \ \textit{almost surely}\,.$$

We begin thus the convergence study of the new opinion dynamics algorithm with choice function

$$\mathbb{P}(u_{t+1}^i = k | \mathcal{F}_t) = P_t^{i,k}.$$

When $f(P) = P$, we have that the ODE (1.11) resolves to a linear ODE of the form

$$\dot{p} = p - D^{-1}Ap = -D^{-1}\Delta p, \tag{A.1}$$

where $\Delta = D - A$ is the graph inward Laplacian.

1. There exists a similar argument when $\beta \ll 1$.

As one can see, the limit Eqn. (A.1) is in fact a linear equation, which means that the limit set for this stochastic approximation algorithm is $\mathcal{K} = \ker \Delta \cap \Delta_K^N$, which means that we must first study $\ker \Delta$ is order to better understand the limit set $\mathcal{K}$.

A great deal of effort was put in order to study spectral properties of the graph Laplacian Δ [59], for which one of the techniques employed was random walks: since $D^{-1}A$ is a stochastic matrix, we can define a random walk on G with state space V and transition probabilities

$$p_{i,j} = D_{i,i}^{-1} A_{i,j},$$

i.e., at each time $t+1$, a random walker[2] that is at node i at time t chooses to go to another node j with probability $p_{i,j} = D_{i,i}^{-1} A_{i,j}$. One can thus apply the theory of Markov chains [238] in order to study the asymptotic properties of this random walk.

Moreover, the stationary states of this random walk are given by the probability row vectors π^T such that $\pi^T = \pi^T D^{-1} A$, i.e., $\pi^T D^{-1} \in \ker \Delta$. This means that studying the stationary states of the Markov chain associated with G sheds light into the limit set $\mathcal{K}$ by transposition. For this study, we need two lemmas: the first one concerns about the form of the eigenvectors (right and left) of the transition matrix $D^{-1}A$, and the second one is a coordinate change which shows that a graph Laplacian always can be factorized into its kernel plus a part with all eigenvalues strictly positive.

Lemma A.1. *Let P be a transition matrix for a finite state Markov chain, i.e., P is a $N \times N$ stochastic matrix, such that there exist C recurrence classes $(C_c, c \leq C)$ for this Markov chain. Then there exist C left invariant and linearly independent probability measures π_c^T and C right eigenvectors 1^c of P such that $\pi_c^T P = \pi_c^T$, $supp(\pi_c) = C_c$, $P1^c = 1^c$ and that, restricted to $\bigcup_c C_c$, we have that $1_i^c = \mathbb{I}_{\{i \in C_c\}}$.*

Proof. Since we have C recurrence classes we can assume, without loss of generality, that

$$P = \begin{pmatrix} P_1 & 0 & 0 & \cdots & 0 \\ 0 & P_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & P_C & 0 \\ T_1 & T_2 & \cdots & T_C & T \end{pmatrix}$$

where $(P_c, c \leq C)$ are irreducible matrices associated with the recurrent classes and T_i are matrices associated with the transient class of the Markov chain.

Since P_c are stochastic matrices, the eigenvalue with maximum value is 1 (it follows from Perron-Frobenius theory). By the Perron-Frobenius theorem, there exist only one eigenvector associated with this eigenvalue, which is the vector 1. The theory of Markov chains gives us a unique left invariant probability π_c for P_c.

Now, if we have a right eigenvector v of P such that $Pv = v$ and $v = (v_1, \cdots, v_C, v_T)$, then $P_c v_c = v_c$ for all $c \leq C$ and $\sum_{c \leq C} T_c v_c + T v_T = v_T$. If $(\mathbb{I} - T)^{-1}$ is well defined, we will have $v_c = 1$ or $v_c = 0$ and $v_T = (\mathbb{I} - T)^{-1} \sum_{c \leq C} T_c v_c$.

We now show that $sp(T) < 1$, which implies that $(\mathbb{I} - T)^{-1}$ is well defined and v_T is uniquely determined by $(T_c, c \leq C)$ and T. Since T is associated with the transient part of the Markov chain,

2. The standard way of defining a random walk in a graph G is to use the outward edges of a node for the random walker. We adopt here the normalized inward edges as transition probabilities in order to study the spectral properties of the normalized inward adjacency matrix $D^{-1}A$.

there exists an index i such that $\sum_j T_{ij} < 1$. Let us assume, without loss of generality, that $i = 1$. Thus $\sum_j (T^2)_{ij} = \sum_j \sum_l T_{il} T_{lj} = \sum_l T_{il} \sum_j T_{lj} = T_{i1} \sum_j T_{1j} + \sum_{l \geq 2} T_{il} \sum_j T_{lj} < T_{i1} + \sum_{l \geq 2} T_{il} \leq 1$, which implies that $sp(T^2) < 1$ and by consequence $sp(T) < 1$. This implies that the eigenvector v such that $Pv = v$ is determined by the components $(v_c, c \leq C)$, which are 1 or 0.

On the other hand, let π^T be a left invariant measure of P such that $\pi^T P = \pi^T$ and $\pi = (\pi_1, \cdots, \pi_C, \pi_T)$. Then $\pi_i^T P_i + \pi_T^T T_i = \pi_i^T$ and $\pi_T^T T = \pi_T^T$. Since $sp(T) < 1$, we have that $\pi_T = 0$ because $\pi_T^T = \pi_T^T T^n$ for every n, which goes to 0. Thus $\pi_i^T P_i = \pi_i^T$ and we know that they are the unique left-invariant probability measures since P_i are irreducible. This shows that the only left invariant probability measures of P are convex combinations of $(0, \cdots, \pi_c, \cdots, 0)$. □

Lemma A.2. *Let C be the algebraic dimension of $\ker D^{-1}\Delta$, i.e., the number of recurrence classes of the Markov chain associated with $D^{-1}A$. There exists a $N \times N$ nonsingular matrix Φ and a $N - C \times N - C$ matrix B such that $\Phi^{-1}(D^{-1}\Delta)\Phi = \begin{pmatrix} 0 & 0 \\ 0 & B \end{pmatrix}$ and B has all generalized eigenvalues strictly positive.*

Proof. Let $(C_c, c \leq C)$ be the recurrence classes of the Markov chain associated with the transition matrix $D^{-1}A$. By lemma A.1 there exists C left invariant probability measures $\pi_c^T, c \leq C$ such that $supp(\pi_c) = C_c$.

Suppose now, by absurd, that there exists a vector v such that $D^{-1}\Delta v \in \ker(D^{-1}\Delta)$, i.e., $D^{-1}\Delta v = x \neq 0$ but $(D^{-1}\Delta)^2 v = 0$.

Let us assume, without loss of generality, that

$$D^{-1}A = \begin{pmatrix} P_1 & 0 & 0 & \cdots & 0 \\ 0 & P_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & P_C & 0 \\ T_1 & T_2 & \cdots & T_C & T \end{pmatrix}$$

where $(P_c, c \leq C)$ are irreducible matrices associated with the recurrent classes and T_i are matrices associated with the transient class of the Markov chain.

Let π^T be an invariant left probability measure of $D^{-1}A$, i.e., $\pi^T D^{-1}A = \pi^T$. Then, $\pi^T x = \pi^T D^{-1}\Delta v = \pi^T (D^{-1}\Delta)^2 v = 0$. Let $x = \sum_n a_n y^n$, where at least one $a_n \neq 0$ and $y^n \in \ker(D^{-1}A)$ and let $x = (x_1, \cdots, x_C, x_T)$, $y^n = (y_1^n, \cdots, y_C^n, y_T^n)$. By lemma A.1 we have that $y_T^n = (\mathbb{I} - T)^{-1} \sum_c T_c y_c^n$, where y_c^n are either 1 or 0.

Since $x \neq 0$ we have two possibilities: $x_c \neq 0$ for at least one $c \leq C$ or $x_c = 0$ for all $c \leq C$ and $x_T \neq 0$. We show that the latter cannot happen: suppose that $x_c = 0$ for all $c \leq C$. This means that $\sum_n a_n y_c^n = 0$ for all $c \leq C$, hence

$$x_T = \sum_n a_n y_T^n = \sum_n a_n (\mathbb{I} - T)^{-1} \sum_{c \leq C} T_c y_c^n = (\mathbb{I} - T)^{-1} \sum_{c \leq C} T_c \sum_n a_n y_c^n = 0$$

and we have that $x = 0$, which is a contradiction.

Suppose then, without loss of generality, that $x_1 \neq 0$ and $\sum_n a_n \neq 0$. Let π^T be an invariant measure such that $supp(\pi) = C_1$. Then $\pi^T x = \sum_n a_n \neq 0$ and we get a contradiction, which shows that the kernel of $D^{-1}\Delta$ can be diagonalized.

Take now $N-C$ linearly independent elements on $\operatorname{im}\Delta$, as $s_1, \cdots, s_{N-C}$. Define the $N\times(N-C)$ matrix S, the $N \times C$ matrix I and the $N \times N$ matrix Φ as

$$S = \begin{pmatrix} | & \cdots & | \\ s_1 & \cdots & s_{N-C} \\ | & \cdots & | \end{pmatrix}, \quad I = \begin{pmatrix} | & \cdots & | \\ 1^1 & \cdots & 1^C \\ | & \cdots & | \end{pmatrix}, \quad \text{and} \quad \Phi = \begin{pmatrix} I \,|\, S \end{pmatrix},$$

where 1^c are N vectors such that $1^c_i = \mathbb{I}_{\{i\in C_c\}}$ on $\bigcup_c C_c$, given by lemma A.1.

We have that Φ is nonsingular since $1^c \in \ker\Delta$, $s_i \in \operatorname{im}\Delta$ and $\mathbb{R}^N \simeq \operatorname{im}\Delta \oplus \ker\Delta$. And since by the Perron-Frobenius theorem applied to $D^{-1}A$, all eigenvalues of $D^{-1}A$ different from 1 have modulus strictly smaller than 1 (which implies that all nonzero eigenvalues of $D^{-1}\Delta$ are strictly positive), we have that

$$D^{-1}\Delta I = 0 \quad \text{and} \quad D^{-1}\Delta S = SB.$$

for some $(N-C)\times(N-C)$ matrix B with all eigenvalues strictly negative. This concludes the proof. □

With both lemmas we are ready to prove the main theorem of this appendix:

Theorem A.1. *Let assumption D.1 be true and $f(P) = P$ be the identity. Then there exists a random variable $P_\infty \in \Delta_K^N$ such that $P_t \to P_\infty$ almost surely.*

Moreover, we also have that for each recurrence class C_c of $D^{-1}A$,

$$\forall k, \quad P_\infty^{i,k} = P_\infty^{j,k} \quad \textit{for all } i,j \in C_c.$$

This means that, if $\ker(D^{-1}\Delta)$ *is unidimensional*[3] *(the Markov chain associated with $D^{-1}A$ has only one recurrence class) we have that P_t reaches a consensus, as in definition 4.*

Proof. Let C be the number of recurrence classes of the Markov chain associated with the stochastic matrix $D^{-1}A$. By lemma A.2 there exists a nonsingular matrix Φ and a $N-C\times N-C$ matrix B such that $\Phi^{-1}(D^{-1}\Delta)\Phi = \begin{pmatrix} 0 & 0 \\ 0 & B \end{pmatrix} = B'$ and B has strictly positive eigenvalues.

Define $Z_t = \Phi^{-1}P_t$, where P_t is the stochastic approximation algorithm (1.9) with $f(P) = P$. We have by Eqn. (1.9) that

$$\begin{aligned} Z_{t+1} = \Phi^{-1}P_{t+1} &= \Phi^{-1}P_t + \frac{1}{t+1}\Big(- \Phi^{-1}D^{-1}\Delta P_t + \Phi^{-1}\beta_{t+1} + \Phi^{-1}\zeta_{t+1}\Big) \\ &= Z_t + \frac{1}{t+1}\Big(- B'Z_t + V_{t+1} + W_{t+1}\Big), \end{aligned}$$

where $W_{t+1} = \Phi^{-1}\zeta_{t+1}$ and $V_{t+1} = \Phi^{-1}\beta_{t+1}$, with $\sum_t \frac{1}{t+1}||V_{t+1}|| < \infty$.

Define the vectors $(Z_t^c)_{c\le C}$ and the $(N-C)\times N$ block matrix Z_t^{N-C} as $Z_t = (Z_t^1, \cdots, Z_t^C, Z_t^{N-C})$, with the same definitions for V and W. By the definition of B', we have that

$$\begin{aligned} Z_{t+1}^c &= Z_t^c + \frac{1}{t+1}(W_{t+1}^c + V_{t+1}^c) \text{ for all } c \le C \\ Z_{t+1}^{N-C} &= Z_t^{N-C} + \frac{1}{t+1}\big(- BZ_t^{N-C} + W_{t+1}^{N-C} + V_{t+1}^{N-C}\big). \end{aligned}$$

3. This is the case if for example the graph G is strongly connected or contains a spanning tree.

Since $\sum_t \frac{1}{(t+1)^2} < \infty$ and ζ_{t+1} is a bounded martingale difference, we have that $\sum_{t' \leq t} \frac{1}{t'+1} W_{t'+1}$ is a L^2-bounded martingale, hence convergent almost surely by Doob's martingale convergence theorem. By the same token, since $\sum_t \frac{1}{t+1} ||V_{t+1}|| < \infty$, we have that $Z^c_{t+1} = Z^c_0 + \sum_{0 \leq t' \leq t} \frac{1}{t'+1} \big(W^c_{t'+1} + V^c_{t'+1}\big)$ converges almost surely to some random variable Z^c_∞.

Since $-B$ has strictly negative eigenvalues, 0 is the only asymptotic stable point of the dynamical system $\dot{z}_t = -Bz_t$. Since $P_\infty \in \Delta_K^N$, we can apply theorem 6.10 of [23] to have that Z_t^{N-C} converges almost surely to 0 when $t \to \infty$.

We have thus $Z_t = (Z_t^1, \cdots, Z_t^c, \cdots, Z_t^C, Z_t^{N-C})$ converges almost surely to $Z_\infty = (Z_\infty^1, \cdots, Z_\infty^C, 0)$, and one has that P_t converges almost surely to $P_\infty = \Phi Z_\infty$ by the continuous mapping theorem. $P_\infty \in \Delta_K^N$ because $P_t \in \Delta_K^N$ for all t and Δ_K^N is closed.

By lemma A.1, the kernel of $D^{-1}\Delta$ is spanned by the vectors 1^c such that restricted to $\bigcup_c C_c$ we have $1^c_i = \mathbb{I}_{\{i \in C_c\}}$. By the proof in lemma A.2, we have that the first C columns of Φ can be assumed, without loss of generality, to be the vectors $(1^c)_{c \leq C}$. This means that for every recurrence class C_c and every $i \in C_c$

$$P_\infty^{i,k} = (\Phi Z_\infty)_{i,k} = (\Phi(Z_\infty^1, \cdots, Z_\infty^C, 0))_{i,k} = \sum_{c \leq C} \Phi_{i,c} Z_\infty^{c,k},$$

where $\Phi_{i,c} = 1^c_i = \mathbb{I}_{\{i \in C_c\}}$. This shows that $P_\infty^{i,k}$ depends only on the recurrence class of i.

Now if $Ker(D^{-1}\Delta)$ is unidimensional, $C = 1$ and we can choose $\Phi = (1, \Phi_{N-1})$ where Φ_{N-1} is a $N \times (N-1)$ matrix. Then $Z_t = (Z_t^1, Z_t^{N-1})$ converges almost surely to $(Z_\infty^1, 0)$ and P_t converges to $P_\infty = \Phi(Z_\infty^1, 0) = (Z_\infty^1, \cdots, Z_\infty^1)$. □

Remark: As one can see, theorem A.1 gives us the convergence of the opinion dynamics algorithm (1.3) to an element of Δ_K^N such that each recurrence class of the stochastic matrix $D^{-1}A$ reaches a consensus. Since each recurrence class of $D^{-1}A$ is a strongly connected component of G, we have created a distributed algorithm that discovers the strongly connected components of G.

APPENDIX B

Estimation of temporal kernel in information diffusion

We derive in this appendix algorithms regarding the estimation of the temporal kernel ϕ defined in chapter 3. These algorithms take two different forms: parametric and nonparametric ones. Due to the large number of nodes of the social networks in question, the nonparametric alternatives are too costly and probably nonviable when dealing with real-life social networks. We present here nevertheless their calculation for the sake of completion.

The two most common employed parametric kernels ϕ are the exponential kernel and the power-law kernel. Exponential kernels [310, 245, 246] are short-range interaction kernels, since they have very light tails, and power-law kernels [71] are long-interaction kernels, since they have heavy tails. Both types of temporal kernels were used in works modeling social networks and social interactions, and their use is entirely up to one's desired modeling considerations - as already discussed in chapter 3 the theory is robust regardless the temporal kernel used.

Remark: It is important to remember that each time we update ϕ, we must recalculate our Hawkes process X_t due to the convolutions $\overline{\phi}_t$. When dealing with exponential kernels, one does not need to use every point in the grid $[0, \tau]$ in order to calculate the exponentials; one may use only up to some fixed length (see [310]). By doing so, one drastically decreases the complexity of this procedure.

In this appendix, for the sake of simplicity, we consider the model of user-user and topic-topic interactions with a single social network and predefined topics, as in chapter 3. The intensity for the cumulative countnumber Hawkes process X_t is given by

$$\lambda_t^{i,k} = \mu^{i,k} + \sum_j \sum_c J_{i,j} B_{c,k} \int_0^{t-} \phi(t-s) dX_s^{j,c},$$

and the log-likelihood is given by (see [74, 240])

$$\mathcal{L} = \sum_{i,k} \left(\int_0^\tau \log(\lambda_t^{i,k}) dX_t^{i,k} - \int_0^\tau \lambda_t^{i,k} dt \right). \tag{B.1}$$

B.1 Parametric estimation of the temporal kernel

There are two standard ways of estimating parametric kernels in Hawkes models, both of them take advantage of the analytic expression of the log-likelihood of X.

The first one is to use numerical optimization methods to find the maximum-likelihood estimator for the kernel parameters by maximizing the log-likelihood (B.1) (as in [15]), and the second one is to derive expectation-maximization methods (see [198, 221, 319]).

Our goal here is to derive expectation-maximization (EM) methods to the kernel parameters, since they are less costly than running numerical optimization schemes, and have the advantage of working nicely with exponential and power-law kernels.

Following Eqn. (B.1), the log-likelihood for this model is thus

$$\begin{aligned}
\mathcal{L} &= \sum_{i,k} \left(\int_0^\tau \log(\lambda_t^{i,k}) dX_t^{i,k} - \int_0^\tau \lambda_t^{i,k} dt \right) \\
&= \sum_{t_l} \log(\lambda_{t_l}^{i_l,k_l}) - \sum_{i,k} \left(\tau \mu^{i,k} + J_{i,i_l} B_{k_l,k} \sum_{t_l} \Phi(\tau - t_l) \right) \\
&= \sum_{t_l} \log(\mu^{i_l,k_l} + \sum_{t_n < t_l} J_{i_l,i_n} B_{k_n,k_l} \phi(t_l - t_n)) - \sum_{i,k} J_{i,i_l} B_{k_l,k} \sum_{t_l} \Phi(\tau - t_l) - \tau \sum_{i,k} \mu^{i,k} \\
&= L_1 - L_2 - \tau \sum_{i,k} \mu^{i,k},
\end{aligned}$$

where

$$\begin{aligned}
L_1 &= \sum_{t_l} \log(\mu^{i_l,k_l} + \sum_{t_n < t_l} J_{i_l,i_n} B_{k_n,k_l} \phi(t_l - t_n)), \\
L_2 &= \sum_{i,k} J_{i,i_l} B_{k_l,k} \sum_{t_l} \Phi(\tau - t_l)
\end{aligned} \tag{B.2}$$

and $\Phi(t) = \int_0^t \phi(s) ds$ is the primitive of the temporal kernel ϕ.

For the EM algorithm, we must insert the parametric kernel ϕ and calculate the maximum likelihood estimator for the kernel parameters. By doing so, we arrive at nonlinear equations for the kernel parameters. One could, at will, use convex optimization methods to calculate this optimal parameter. We adopt a different path: we approximate the nonlinear parts of the derivative of the log-likelihood in order to get analytic and simple updates for the parameters.

We perform this linear approximation is order to decrease the complexity of the algorithm, since the log-likelihood has a quadratic complexity on the jumps of X, due to the double sum $\sum_{t_l} \sum_{t_n < t_l}$.

Algorithm 4 - EM estimation procedure

1: **Input:** jumps times t_l, Hawkes parameters J, B, μ, initial condition (ω^0 for exponential, a^0 and b^0 for power-law)
2: **while** Temporal kernel parameters have not converged **do**
3: Calculate branching variables p_l^0 and $(p_l^n)_{n<l}$ with parameters of previous step
4: Calculate parameters EM update equation with new branching variables (Eqn. (B.3) for the exponential and Eqn. (B.4) for the power-law - calculate a and b in a cyclic manner)
5: **end while**
6: **Output:** Kernel parameters (ω for exponential, a and b for power-law)

B.1.1 Expectation-maximization algorithm for an exponential kernel

We start the parametric estimation of the temporal kernel with an exponential kernel of the form $\phi(t) = \omega e^{-\omega t}.\mathbb{I}_{t>0}$, with $\omega > 0$. Let t_l be the jumps of the Hawkes process, and let i_l be the user that broadcasted content k_l at time t_l.

From Eqn. (B.2) we have that the log-likelihood of X with an exponential temporal kernel takes the form

$$L_1 = \sum_{t_l} \log(\mu^{i_l,k_l} + \sum_{t_n<t_l} J_{i_l,i_n} B_{k_n,k_l} \omega e^{-\omega(t_l-t_n)}),$$
$$L_2 = \sum_{i,k} J_{i,i_l} B_{k_l,k} \sum_{t_l} (1 - e^{-\omega(\tau-t_l)}).$$

Using the concavity of the logarithm function, we have that

$$\begin{aligned} L_1 &\geq \sum_{t_l} \left(\sum_{t_n<t_l} \left(p_l^n \log(J_{i_l,i_n} B_{k_n,k_l} \omega e^{-\omega(t_l-t_n)}) - p_l^n \log p_l^n\right) + p_l^0 \log(\mu^{i_l,k_l}) - p_l^0 \log p_l^0 \right) \\ &= \sum_{t_l} \left(\sum_{t_n<t_l} \left(p_l^n \log(J_{i_l,i_n} B_{k_n,k_l} \omega) - p_l^n \omega(t_l - t_n) - \mathcal{E}(p_l^n)\right) + \mathcal{H}(p_l^0) \right) \\ &= L_1^p \end{aligned}$$

where $\mathcal{E}(x) = x \log x$, $\mathcal{H}^l(p_l^0) = p_l^0 \log(\mu^{i_l,k_l}) - \mathcal{E}(p_l^0)$ and the nonnegative branching variables p_l^0, p_l^n satisfy $p_l^0 + \sum_{n<l} p_l^n = 1$.

Maximizing L_1^p with respect to the branch variables, under the constraint $p_l^0 + \sum_{n<l} p_l^n = 1$, gives us

$$p_l^0 = \frac{\mu^{i_l,k_l}}{\mu^{i_l,k_l} + \sum_{t_n<t_l} J_{i_l,i_n} B_{k_n,k_l} \omega e^{-\omega(t_l-t_n)}} \text{ and } p_l^n = \frac{J_{i_l,i_n} B_{k_n,k_l} \omega e^{-\omega(t_l-t_n)}}{\mu^{i_l,k_l} + \sum_{t_n<t_l} J_{i_l,i_n} B_{k_n,k_l} \omega e^{-\omega(t_l-t_n)}}.$$

Now, maximizing $\mathcal{L}^p = L_1^p - L_2$ with respect to ω gives us

$$\partial_\omega \mathcal{L}^p = \sum_{t_l} \left(\sum_{t_n<t_l} p_l^n \left(\frac{J_{i_l,i_n} B_{k_n,k_l}}{\omega} - (t_l - t_n)\right) - \sum_{i,k} J_{i,i_l} B_{k_l,k} (\tau - t_l) e^{-\omega(\tau-t_l)} \right) = 0.$$

Let ω^t be the output of the EM algorithm at step t. By approximating $e^{-\omega^{t+1}(\tau-t_l)}$, the exponential at time $t+1$, by $e^{-\omega^t(\tau-t_l)}$, the exponential at time t, we get a linear update for ω^{t+1} as

$$\omega^{t+1} = \frac{\sum_{t_l} \sum_{t_n<t_l} p_l^n J_{i_l,i_n} B_{k_n,k_l}}{\sum_{t_l} \left(\sum_{t_n<t_l} p_l^n (t_l - t_n) + \sum_{i,k} J_{i,i_l} B_{k_l,k} (\tau - t_l) e^{-\omega^t(\tau-t_l)} \right)}. \tag{B.3}$$

B.1.2 Expectation-maximization algorithm for a power-law kernel

We now derive an EM algorithm for a power law kernel of the form $\phi(t) = b(a+t)^{-(b+1)}$, for $a > 0$ and $b \neq 0$. From Eqn. (B.2) we have that the log-likelihood of X with a power-law temporal kernel takes the form

$$L_1 = \sum_{t_l} \log(\mu^{i_l,k_l} + \sum_{t_n<t_l} J_{i_l,i_n} B_{k_n,k_l} b(a + t_l - t_n)^{-(b+1)}),$$
$$L_2 = \sum_{i,k} J_{i,i_l} B_{k_l,k} \sum_{t_l} (a^{-b} - (a + \tau - t_l)^{-b}).$$

Using again the concavity of the logarithm, we have the lower bound for L_1,

$$\begin{aligned} L_1 &\geq \sum_{t_l} \Bigg(\sum_{t_n<t_l} \big(p_l^n \log(J_{i_l,i_n} B_{k_n,k_l} b) - p_l^n (b+1) \log(a + t_l - t_n) - \mathcal{E}(p_l^n) \big) + \mathcal{H}(p_l^0) \Bigg) \\ &= L_1^p. \end{aligned}$$

Thus, maximizing this bound for p_l^0 and p_l^n under the constraint $p_l^0 + \sum_{n<l} p_l^n = 1$ gives us

$$p_l^0 = \frac{\mu^{i_l,k_l}}{\mu^{i_l,k_l} + \sum_{t_n<t_l} J_{i_l,i_n} B_{k_n,k_l} b (a + t_l - t_n)^{-(b+1)}}$$

and

$$p_l^n = \frac{J_{i_l,i_n} B_{k_n,k_l} b (a + t_l - t_n)^{-(b+1)}}{\mu^{i_l,k_l} + \sum_{t_n<t_l} J_{i_l,i_n} B_{k_n,k_l} b (a + t_l - t_n)^{-(b+1)}}.$$

And maximizing $\mathcal{L}^p = L_1^p - L_2$ with respect to a and b gives

$$\begin{aligned} \partial_a \mathcal{L} &= \sum_{t_l} \Bigg(- \sum_{t_n<t_l} \frac{p_l^n (b+1)}{a + t_l - t_n} + b \sum_{i,k} J_{i,i_l} B_{k_l,k} (a^{-(b+1)} - (a + \tau - t_l)^{-(b+1)}) \Bigg) \\ \partial_b \mathcal{L} &= \sum_{t_l} \Bigg(\sum_{t_n<t_l} \big(\frac{p_l^n J_{i_l,i_n} B_{k_n,k_l}}{b} - p_l^n \log(a + t_l - t_n) \big) \\ &\quad + \sum_{i,k} J_{i,i_l} B_{k_l,k} \big(a^{-b} \log a - (a + \tau - t_l)^{-b} \log(a + \tau - t_l) \big) \Bigg). \end{aligned}$$

Again, let a^t and b^t be the updates of a and b at step t. Approximating $\big(a^{-b^{t+1}} \log a - (a + \tau - t_l)^{-b^{t+1}} \log(a+\tau-t_l)\big)$ by $\big(a^{-b^t} \log a - (a+\tau-t_l)^{-b^t} \log(a+\tau-t_l)\big)$ and $\frac{p_l^n (b+1)}{a^{t+1}+t_l-t_n}$, $(a^{t+1}+\tau-t_l)^{-(b+1)}$ by $\frac{p_l^n (b+1)}{a^t+t_l-t_n}$, $(a^t + \tau - t_l)^{-(b+1)}$, respectively, we get

$$a^{t+1} = \left(\frac{b \sum_{t_l} \sum_{i,k} J_{i,i_l} B_{k_l,k}}{\sum_{t_l} \left(\sum_{t_n<t_l} \frac{p_l^n (b+1)}{a^t + t_l - t_n} + b \sum_{i,k} J_{i,i_l} B_{k_l,k} (a^t + \tau - t_l)^{-(b+1)} \right)} \right)^{\frac{1}{b+1}} \tag{B.4}$$

and

$$b^{t+1} = \frac{\sum_{t_l} \sum_{t_n<t_l} p_l^n J_{i_l,i_n} B_{k_n,k_l}}{\sum_{t_l} \left(\mathcal{B}_l^1 + \sum_{i,k} J_{i,i_l} B_{k_l,k} \mathcal{B}_l^2 \right)},$$

where $\mathcal{B}_l^1 = \sum_{t_n<t_l} p_l^n \log(a + t_l - t_n)$ and $\mathcal{B}_l^2 = \big((a + \tau - t_l)^{-b^t} \log(a + \tau - t_l) - a^{-b^t} \log a \big)$.

B.2 Nonparametric estimation of the temporal kernel

As next step of the temporal kernel estimation, we describe a nonparametric procedure, following Bacry and Muzy [17] and using an intensity of the form

$$\lambda_t^{i,k} = \mu^{i,k} + \sum_{j,c} J_{i,j} B_{c,k} \int_{-\infty}^{t} \phi(t-s) dX_s^{j,c}.$$

We assume that the Hawkes process is in the stable regime (see chapter 4). In order for X to be in the stable regime, we must have (see lemma 24)

$$sp(B)sp(J) < \frac{1}{||\phi||_1}.$$

Since in the stable regime the Hawkes process X has stationary increments (see [140, 44]), we define $\Lambda^{i,k} = \mathbb{E}[\lambda_t^{i,k}]$ to be the expectation of the intensity, which satisfies

$$\begin{aligned}\Lambda^{i,k} = \mathbb{E}[\lambda_t^{i,k}] &= \mu^{i,k} + \sum_{j,c} J_{i,j} B_{c,k} \int_{-\infty}^{t} \phi(t-s)\mathbb{E}[dX_s^{j,c}] \\ &= \mu^{i,k} + \sum_{j,c} J_{i,j} B_{c,k} \int_{-\infty}^{t} \phi(t-s)\Lambda^{j,c},\end{aligned}$$

which in matrix form becomes

$$\Lambda = \mu + ||\phi||_1 J\Lambda B. \tag{B.5}$$

By taking the vectorization of Eqn. (B.5), we get

$$v(\Lambda) = (\mathbb{I} - ||\phi||_1 (B^T \otimes J))^{-1} v(\mu).$$

It has been shown in [17] that the first and second order statistics of a Hawkes process completely characterizes it, i.e., we only need the expectation and covariance functions to determine its structure. Since Hawkes processes are orderly processes (they almost surely do not possess more than one jump at the same time) with jumps of size 1, we can use the conditional expectation of jumps instead of the covariance function, defined by

$$G^{(i,j),(c,k)}(t)dt = \mathbb{E}[dX_t^{i,k} | dX_0^{j,c} = 1] - \delta_{(i,k),(j,c)}\delta(t) - \Lambda^{i,k}dt,$$

where $\delta_{l,n}$ is the Kronecker delta, i.e., $\delta_{l,n} = 1$ if $l = n$ and 0 otherwise.

The conditional expectation can be seen as a function $G : \mathbb{R} \to \mathcal{M}_{NK\times NK}(\mathbb{R})$ by using the indices $(i,j),(c,k)$ as in $(B^T \otimes J)$. Defining $\Phi(t)$ to be the nonparametric $NK \times NK$ matrix composed by the kernels of the Hawkes process X, as

$$v(\lambda_t) = v(\mu) + (\Phi * dX)_t,$$

we have by proposition 3 of [17] that Φ satisfies the N^2K^2 dimensional Wiener-Hopf system

$$G(t) = \Phi(t) + \Phi * G(t), \quad t > 0, \tag{B.6}$$

which admits a unique solution, i.e., it completely characterizes the temporal kernel Φ. Bacry and Muzzy developed in [17] a nonparametric estimation procedure for the whole kernel matrix Φ, which does not take into account its partial parametric form $\Phi(t) = \phi(t)(B^t \otimes J)$.

In our case, we do not have for each time t the full N^2K^2 degrees of freedom from $\Phi(t)$, since we have that the kernels must satisfy the constraint $\Phi(t) = \phi(t)(B^t \otimes J)$. That being said, our nonparametric estimation of $\phi(t)$ is thus a problem of finding the function $\phi(t)$ that best approximates the true kernels $\Phi(t)$.

In order to do so, we derive from the Wiener-Hopf system (B.6) the equations satisfied by G, as in [17], which gives us an equality between the tensor G and the kernel ϕ. Then, we discretize $[0,\tau]$

into bins to achieve a linear system linking both quantities, and apply a nonnegative least squares estimation procedure to determine the best ϕ in each time bin.

By Eqn. (30) in [17], we have that G satisfies

$$\begin{aligned} G^{(i,j),(c,k)}(t) &= \phi^{(i,j),(c,k)}(t) + \sum_{l,p}\Bigg(\int_{s>0}\phi^{(i,l),(p,k)}(t-s)G^{(l,j),(c,p)}(s)ds \\ &+ \frac{\Lambda^{l,p}}{\Lambda^{j,c}}\int_{s<0}\phi^{(i,l),(p,k)}(t-s)G^{(j,l),(p,c)}(-s)ds\Bigg) \\ &= J_{i,j}B_{c,k}\phi(t) + \sum_{l,p}J_{i,l}B_{p,k}\int_{s>0}\left(\phi(t-s)G^{(l,j),(c,p)}(s) + \frac{\Lambda^{l,p}}{\Lambda^{j,c}}\phi(t+s)G^{(j,l),(p,c)}(s)\right)ds \end{aligned} \tag{B.7}$$

For simplicity, take a size T quadrature[1] grid $[0, h, 2h, \cdots, sh, \cdots, (T-1)h]$ of $[0, \tau]$, with timestep h, and let $l = i + N(k-1)$ and $n = j + N(c-1)$, where l and n are then indices for the vectorization of Λ. Define, with abuse of notation, the $NK \times NK$ tensor Ψ, the T vector ϕ and the $NK \times NK \times T$ tensor G such that

$$\Psi = B^T \otimes J, \quad \phi_s = \phi((s-1)h), \quad G_{l,n,s} = G^{l,n}((s-1)h),$$

Using the quadrature grid, Eqn. (B.7) becomes[2]

$$G_{l,n,t} = \Psi_{l,n}\phi_t + h\sum_{s\geq 1}\sum_{d}\Psi_{l,d}\left(G_{d,n,s}\phi_{t-s} + \frac{\Lambda_d}{\Lambda_n}G_{n,d,s}\phi_{t+s}\right), \tag{B.8}$$

where $\phi_{t-s} = 0$ if $s \geq t$ from the fact that ϕ is a causal kernel.

Defining the $NK \times NK \times T \times T$ tensor η, such that

$$\begin{cases} \eta_{l,n,t,s} &= \Psi_{l,n} \text{ if } s = t \\ \eta_{l,n,t,s} &= h\sum_d \Psi_{l,d}G_{d,n,t-s} \text{ if } s < t \\ \eta_{l,n,t,s} &= h\sum_d \Psi_{l,d}\frac{\Lambda^d}{\Lambda^n}G_{n,d,s-t} \text{ if } s > t, \end{cases}$$

we have $G = \eta \times_4 \phi^T$, where $\eta \times_4 \phi^T$ is the mode-1 product of η and ϕ^T (see [170]). And defining the $NK \times NK \times T \times T$ tensor ρ by

$$\begin{cases} \rho_{l,n,t,s} &= \delta_{l,n} \text{ if } s = t \\ \rho_{l,n,t,s} &= hG_{d,n,t-s} \text{ if } s < t \\ \rho_{l,n,t,s} &= h\frac{\Lambda^d}{\Lambda^n}G_{n,d,s-t} \text{ if } s > t, \end{cases}$$

we have that $\eta = \rho \times_1 \Psi$.

We easily have by Eqn. (B.8) then

$$G = \eta \times_4 \phi^T = \rho \times_1 \Psi \times_4 \phi^T.$$

We can thus apply a nonnegative Tucker decomposition with quadratic cost function to find the best nonnegative ϕ, following [170], which gives us

$$\phi^T \leftarrow \phi^T \odot \frac{[G \times_1 (B \otimes J^T)]_{(4)}\rho_{(4)}^T}{\phi^T[\rho \times_1 (BB^T \otimes J^TJ)]_{(4)}\rho_{(4)}^T},$$

1. In [17], the estimation procedure is based on the Nyström method [54] using Gaussian quadrature.
2. One can notice that Eqn. (B.7) has a discontinuity in $t = 0$, which is also observed in Eqn. (B.8). One way to deal with this problem is to estimate directly the integrals, as in Eqn. (43) of [17].

with $G_{(4)}$ the mode-4 matricization of the tensor G (see [170]). Taking the transpose amounts to the following multiplicative updates

$$\phi \leftarrow \phi \odot \frac{\rho_{(4)}[G \times_1 (B \otimes J^T)]_{(4)}^T}{\rho_{(4)}[\rho \times_1 (BB^T \otimes J^T J)]_{(4)}^T \phi}, \tag{B.9}$$

which converges to the unique minimum, since the quadratic cost is convex in ϕ.

Remark: Another estimation technique to solve nonnegative Tucker decomposition problems in the alternating least squares (ALS) method [62, 251, 250], in which one replaces the cyclic multiplicative updates to alternating projected gradient descents.

One can see that, nevertheless, the complexity for this update is high, due to the matrix product from the matricization of the tensors. This approach is useful, however, if one wants to estimate separately the matrix $\Psi = B^T \otimes I$.

It is also important to notice that the NTD updates are matrix products and entrywise divisions and multiplications, so they can be performed in a distributed fashion. Moreover, one may use the structure in Ψ and ρ is order to decrease the complexity of the mode-4 and mode-1 products, and the mode-4 matricizations.

The simpler approach is to use the fact that

$$G = \eta \times_4 \phi^T \;\Leftrightarrow\; G_{(4)}^T = \eta_{(4)}^T \phi,$$

which can be seen as the matrix multiplication, and perform a nonnegative least squares estimation in order to find ϕ.

Algorithm 5 - Nonparametric estimation procedure

1: **Input:** jumps of Hawkes process dX, Hawkes parameters J, B, μ, discretization timestep h
2: Estimate Λ by $\Lambda^{i,k} \simeq \frac{X_\tau^{i,k}}{\tau}$
3: Discretize $[0, \tau]$ as $[0, h, 2h, \cdots, (T-1)h]$
4: Estimate $G_{l,n,s}$ using empirical averages
5: Calculate matrices Ψ, η and ρ (and the possible mode-m products and matricizations)
6: Estimate ϕ using nonnegative least squares (or the NTD multiplicative updates (B.9) until convergence)
7: **Output:** Kernel function $\phi((s-1)h),\ s \in \{1, \cdots, T\}$

APPENDIX C

Modified estimation of topic models

C.1 Introduction

Chapter 3 is concerned with the development and study of our Hawkes-based information diffusion framework. One of the biggest advantages of this framework is that one can create and adapt different models at ease, such as the models derived in subsections 3.2.1, 3.2.2, 3.2.3, 3.2.4 and 3.2.5

From all abovementioned models of information diffusion, one of them is in need of a tool that does not stem from point processes and nonnegative tensor decompositions in order to be properly analyzed: the "fuzzy" diffusion model of subsection 3.2.3. The tool necessary for the adequate treatment of "fuzzy" diffusions is *topic models* [34, 265, 145].

Topic modeling is a subfield of machine learning and natural language processing. A topic model is a statistical model for discovering compound "topics", composed of multiple ideas, that occur in documents. Intuitively speaking, given that a document is about a particular topic, one would expect particular words to appear in the document more or less frequently: "politics" or "vote" would probably appear more in a political-oriented document. A document which is 80% about politics and 20% about economics would have on average four times more words linked to politics than it has about economics. A topic model captures this intuition in a mathematical framework, which allows examining documents and discovering, based on the statistics of the words contained in each of them, what would be the relevant topics and their balance.

In this appendix we take advantage of the *probabilistic latent semantic analysis* (PLSA) framework [152, 222], which uses statistical techniques to study relationships between words in documents. One of the ideas used in PLSA is that a document is a mixture decomposition stemming from a latent class model, i.e., documents are (probabilistic) combinations of "yet undiscovered topics". Among topic models from PLSA, two are particularly useful in our Hawkes-based information diffusion framework: *latent Dirichlet allocation* [34] and *author-topic model* [265].

Latent Dirichlet allocation (LDA) is a generative probabilistic model for discrete data, such as text, developed in [34] by Blei *et al.* It is based on a hierarchical Bayesian model, in which every document is an independent finite mixture over K topics. Each topic is then represented as a finite mixture over a representative set of topic probabilities. In other terms, documents are probabilities over the K topics and these probabilities will be related to the words encountered in each document.

In the LDA, words have topic assignments that are document-dependent and independent from each other, and documents are independent mixtures of the K latent topics, which are retrieved from its words. These topics are thus defined as distributions over words, and this mechanism defines a hierarchical Bayesian model.

Author-topic model (ATM) is also a generative probabilistic model for discrete data, similar to

the LDA. The big difference between both models is that in the ATM, the topic variables assigned to words in documents are not document-oriented, but author-oriented, i.e., in the LDA the words of a document were intrinsically related to the document in question, whereas in the ATM words are related to the authors of the document.

The motivation for using the ATM instead of the LDA is the fact that in the LDA, messages are a mixture of topics, independent of other messages. This means that we do not take into consideration the authors inclination to post messages on their topics of expertise or interest. Think for example about Barack Obama: he is more likely to post messages on Twitter about topics related to economics or politics than topics related to fashion or sports. The ATM takes that individuality into consideration when discovering the latent topics, as opposed to the LDA which assumes that each tweet from Obama is independent from the others.

A potential weakness of the ATM is that it does not allow any particular aspect of a document. The document is thus generated only by a mixture of the authors' topic distributions. The LDA model on the other hand is in a sense its complete opposite - it allows each document to have its own document-specific topic mixture. One could also provide less "extreme" topic models that lie between these two, allowing a flexible treatment of the broadcasted messages.

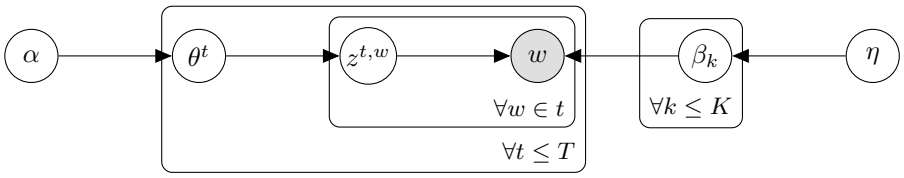

Figure C.1: Graphical model for latent Dirichlet allocation [34].

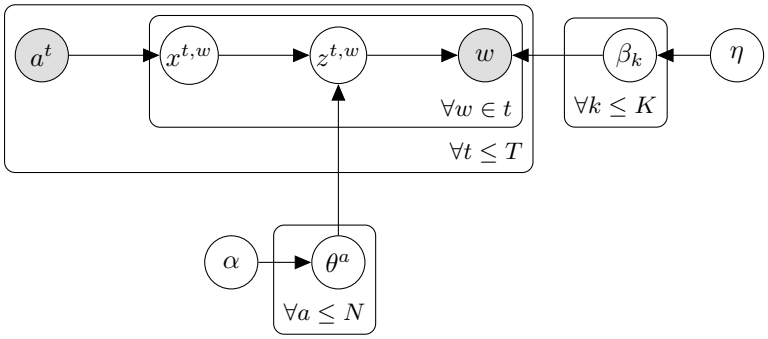

Figure C.2: Graphical model for author-topic language model [265].

The use of the LDA and ATM topic models in this thesis is not restrictive at all and different topic models are in fact plausible. We have chosen these particular two topic models for their simplicity, which makes them base models for a series of extensions and variants [271, 280, 299].

Since we are not dealing with a standard topic model problem, we will exchange the nomenclature with our own, in order to avoid confusion:

- Documents become **messages** that are broadcasted.
- Messages are a mixture of **topics** or *contents*.

- Authors are the **users** in our social networks.
- As in LDA and ATM, a **word** is the basic unit of data, being defined as an entry of a vocabulary (dictionary) of size W. A word relative to the i^{th} entry of the vocabulary is a W-dimensional vector w such that $w_i = 1$ and $w_j = 0$ for $j \neq i$.

C.2 Generative procedures for latent Dirichlet allocation and author-topic model

C.2.1 Latent Dirichlet allocation

The latent Dirichlet allocation is a generative probabilistic topic model following the generative process (see Figure C.1):

1- Choose the K topic distributions over the W words, $\beta_k,\ k \in \{1, 2, \cdots, K\}$, such that[1]

$$\beta_k \sim Dirichlet(\eta).$$

2- For each message $t \in \{1, 2, \cdots, T\}$ we choose the message topic distribution

$$\theta^t \sim Dirichlet(\alpha).$$

2.1- For each word w in the message t we get the topic $z^{t,w} \in \mathbb{R}_+^K$ from the discrete[2] distribution θ^t, such as

$$\mathbb{P}(z_k^{t,w} = 1|\theta^t) = \theta_k^t. \tag{C.1}$$

2.2- Given the topic $z^{t,w}$ we choose the word topic from the $K \times W$ matrix topic distribution β as

$$\beta_{kj} = \mathbb{P}(w_j = 1|z_k^{t,w} = 1).$$

C.2.2 Author-topic model

The author-topic model is a generative probabilistic topic model following the generative process (see Figure C.2):

1- Choose the K topic distributions over the W words, $\beta_k,\ k \in \{1, 2, \cdots, K\}$, such that

$$\beta_k \sim Dirichlet(\eta).$$

2- For each user/author in V, choose the author-topic distributions, $\theta^a,\ a \in \{1, 2, \cdots, N\}$, such that

$$\theta^a \sim Dirichlet(\alpha).$$

3- For each message $t \in \{1, 2, \cdots, T\}$ with authors $a_t \subset \{1, 2, \cdots, N\}$

1. A random variable β follows a Dirichlet distribution with parameter η (such that $\eta_j > 0$ for all j), denoted by $\beta \sim Dirichlet(\eta)$, if β has probability density $p(x) = \frac{\prod_j \Gamma(\eta_j)}{\Gamma(\sum_j \eta_j)} \prod_j x_j^{\eta_j - 1}$ for x such that $x_j \geq 0$ and $\sum_j x_j = 1$.
2. A random variable $z \in \{1, 2, \cdots, K\}$ follows a discrete distribution with parameter θ (such that $\theta_k \geq 0$ and $\sum_k \theta_k = 1$), denoted by $z \sim Discrete(\theta)$, if $\mathbb{P}(z = k) = \theta_k$.

3.2- For each word w in the message t, we choose the author $x^{t,w} \in a^t = \{a^{i_1}, \cdots, a^{i_t}\}$ corresponding to the word w in a uniform fashion

$$\mathbb{P}(x^{t,w} = a^{i_n}) = \frac{1}{\sharp a^t}.$$

3.2- Given the author $x^{t,w}$ of word w in message t, we get the topic $z^{t,w} \in \mathbb{R}_+^K$ from the discrete distribution $\theta^{x^{t,w}}$, such as

$$\mathbb{P}(z_k^{t,w} = 1 | \theta^{x^{t,w}}) = \theta_k^{x^{t,w}}. \tag{C.2}$$

3.3- Given the topic $z^{t,w}$, we choose the word topic from the $K \times W$ matrix topic distribution β as

$$\beta_{kj} = \mathbb{P}(w_j = 1 | z_k^{t,w} = 1). \tag{C.3}$$

Remark: The random variable $x^{t,w}$ represents the author associated with the word w in document t, sampled uniformly from $a_t \subset \{1, 2, \cdots, N\}$. In our case, $\sharp a^t = 1$ for all messages $t \leq T$, i.e., each message has only one author and this author is the user $i_t \in \{1, 2, \cdots, N\}$ that broadcasted the message at time t_t.

Remark: In both topic models, we have two Dirichlet hyperparameters η and α responsible to "smooth" the topic distributions β, θ and give them a predetermined shape. A great deal of importance is given to them, with detailed and thorough discussions [144, 12, 296] from theoretical and practical points of view.

C.3 Topic model parameters estimation

There are several ways of estimating the LDA topic model parameters, such as variational methods, Gibbs sampling, expectation-propagation, among others: for example, Blei *et al.* [34] develop a variational Bayes algorithm, Hoffman *et al.* [151] develop an online variational Bayes algorithm and [47, 77, 128] illustrate a Gibbs sampler procedure for the estimation of the beforementioned parameters.

We concentrate here on the two most common ways of estimating the topic model parameters θ, z and β: *variational Bayes methods* (see [34, 151, 145]) and *Gibbs sampling methods* (see [47, 77, 128]).

As already mentioned, we place ourselves on the model in subsection 3.2.3, which serves as an example for every "fuzzy" diffusion model implemented in our information diffusion framework of chapter 3.

We take advantage of the relationship between the topic model and the information diffusion cascades generated by the Hawkes process X, which stems from the presence of the random variables Z^{t_s} in the intensity, rewritten here

$$\lambda_t^i = \mu_i + \sum_{j,c,k} J_{i,j} B_{c,k} b_{i,k} \int_0^{t-} \phi(t-s) Z_c^s dX_s^j.$$

We use then the relationship between the topic model and the Hawkes process X to derive a more data-driven methodology for the topic model parameters estimation. Other examples of such methodology are [310], where the authors introduce a simple topic model to detect the mutation of memes in social networks, and [200], where the authors introduce the LDA topic model to identify and label search tasks and queries.

C.3.1 Modified Collapsed Gibbs sampler

We start our estimation methodology with the Gibbs sampler (see [114, 28, 115] for a more detailed discussion). Gibbs sampling is a member of the Markov-chain Monte Carlo (MCMC) framework (see [28, 115]), and it is a particular instance of the Metropolis-Hastings algorithm. In Bayesian estimation, MCMC algorithms aim to construct a Markov chain that has the posterior distribution as its unique stationary distribution.

In both topic models, one wants to sample from the posterior of z, i.e., $\mathbb{P}(z|w,\alpha,\eta)$, however this expression is unknown. On the other hand, since the discrete distribution and the Dirichlet distribution are conjugate [257, 113], one can analytically integrate θ and β out of the posterior. This gives rise to a simple and efficient estimation method called *collapsed Gibbs sampling* [209] (where the discrete random variable θ and the Dirichlet random variable β are "collapsed" from the posterior distribution).

Remark: Since we have only one user broadcasting each message in our framework, we have by definition $\sharp a^t = 1$ for every message t, i.e., the ensemble of authors for each message in the ATM is always unitary, the posterior distribution $\mathbb{P}(z,a,\theta,\beta|w,x,\alpha,\eta)$ for the ATM can be simplified as $\mathbb{P}(z,\theta,\beta|w,x,\alpha,\eta)$.

From now on, let $z^{s,i}$ be the topic of word w^i, belonging to message[3] s, and $z^{-(s,i)}$ be the topics of all other words except the word w^i. Let also $x^{s,i}$ be the author of word w^i, belonging to message s, and $x^{-(s,i)}$ be the authors of all other words except the word w^i.

The standard collapsed Gibbs sampling method for sampling z works as follows: one wants to sample from the posterior $\mathbb{P}(z,|w,\alpha,\eta)$, however this posterior is unknown due to dependencies between z. One can, on the other hand, calculate analytically $\mathbb{P}(z^{s,i},|w,z^{-(s,i)},\alpha,\eta)$ for both LDA and ATM by collapsing the latent random variables θ and β. Moreover, one can show that to sample from $\mathbb{P}(z,|w,\alpha,\eta)$, it suffices sampling from $\mathbb{P}(z^{s,i},|w,z^{-(s,i)},\alpha,\eta)$ in a cyclical way, since by Bayes rule

$$\begin{aligned}\mathbb{P}(z,|w,\alpha,\eta) &= \mathbb{P}(z^{s,i},z^{-(s,i)}|w,\alpha,\eta) = \mathbb{P}(z^{s,i}|w,z^{-(s,i)},\alpha,\eta)\mathbb{P}(z^{-(s,i)}|w,\alpha,\eta)\\ &\propto \mathbb{P}(z^{s,i}|w,z^{-(s,i)},\alpha,\eta),\end{aligned}$$

i.e., sampling from $\mathbb{P}(z,|w,\alpha,\eta)$ and from $\mathbb{P}(z^{s,i}|w,z^{-(s,i)},\alpha,\eta)$ is equivalent since they are proportional.

We can use the analytic form of the standard Gibbs sampler to derive modified collapsed Gibbs sampling method for the LDA and ATM, in two different parts. The methodology is the same for both of them, differing only on the notation used (the ATM possesses authors, which change the equation for the posterior distribution). The difference between our modified version and the standard version of Gibbs sampling stems from the introduction of the point process X into the estimation, using Bayes rule.

Remark: We do not discuss here about the practical implementation of a Gibbs sampler, since it has already been largely discussed in the literature [115]. We only present the modified sampling equations for the sampler.

Remark: The fact that the Gibbs sampling is quite simple for the LDA and ATM makes another good reason to why choose these models as base models to our information diffusion framework.

3. From now on, we denote message s the message broadcasted at time t_s.

C.3.1.1 Latent Dirichlet allocation

We have from [47, 77, 128] that the sampling probabilities for $z^{i,s}$ can be calculated analytically using $z^{-(s,i)}$, w, α and η as

$$\mathbb{P}(z^{s,i} = k | z^{-(s,i)}, w, \alpha, \eta) \propto \frac{n_{-i,k}^{(w^i)} + \eta_{w^i}}{n_{-i,k}^{(\cdot)} + \sum_j \eta_j} . \frac{n_{-i,k}^{(s)} + \alpha_k}{n_{-i,\cdot}^{(s)} + \sum_{k'} \alpha_{k'}}, \tag{C.4}$$

where

- $n_{-i,k}^{(w^i)}$ is the number of instances of word w^i assigned to topic k, in exception of word w^i in message s,
- $n_{-i,k}^{(\cdot)}$ is the total number of words, in exception of word w^i in message s, that are assigned to topic k,
- $n_{-i,k}^{(s)}$ is the number of words in message s assigned to topic k, in exception of word w^i,
- $n_{-i,\cdot}^{(s)}$ is the total number of words in message s, in exception of word w^i.

Now we introduce the Hawkes process X_t: Let X be the instances of the point process X_t and Z be the empirical topic proportions of messages, given by Eqn. (3.2). Since the intensity λ_t depends only on z through Z, we have by Bayes rule

$$\begin{aligned}
&\mathbb{P}(z^{s,i} | z^{-(s,i)}, w, X, \alpha, \eta) \propto \mathbb{P}(X | z^{s,i}, z^{-(s,i)}, w, \alpha, \eta) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, w, \alpha, \eta) \\
&= \mathbb{P}(X | z^{s,i}, z^{-(s,i)}) \mathbb{P}(z^{s,i} | z^{-(s,i)}, w, \alpha, \eta) = \mathbb{P}(X|Z) \mathbb{P}(z^{s,i} | z^{-(s,i)}, w, \alpha, \eta) \\
&= L(X|Z) \mathbb{P}(z^{s,i} | z^{-(s,i)}, w, \alpha, \eta),
\end{aligned} \tag{C.5}$$

where $L(X|Z)$ is the conditional likelihood of X given Z, as (see [74, 240])

$$L(X|Z) = \Big[\prod_{n=1}^{X(\tau)} \lambda_{t_n}^{i_n} \Big] \exp(- \sum_i \int_0^\tau \lambda_u^i du), \tag{C.6}$$

where i_n is the user that broadcasted the message at time t_n and $X(\tau)$ is the total number of jumps of X in $[0, \tau]$, i.e., the total number of messages broadcasted in $[0, \tau]$.

Let t_s be the time of broadcast of the message s containing word w^i and i_s be the user that broadcasted the message at time t_s. Looking at the likelihood $L(X|Z)$ more closely, one can see some terms that do not depend on $z^{s,i}$, and are casted out during the normalization process; these are the terms not containing Z^{t_s}. Thus one can replace $L(X|Z)$ by

$$\begin{aligned}
L^{t_s}(X|Z) &= \Big[\prod_{t_n > t_s}^{X(\tau)} \lambda_{t_n}^{i_n} \Big] \exp\Big(- \int_0^\tau \sum_{j,c,k} J_{j,i_s} B_{c,k} Z_c^{t_s} b_{j,k} \phi(t - t_s) dt \Big) \\
&= \Big[\prod_{t_n > t_s}^{X(\tau)} \lambda_{t_n}^{i_n} \Big] \exp\Big(- \sum_{j,c,k} J_{j,i_s} B_{c,k} Z_c^{t_s} b_{j,k} \Phi(\tau - t_s) \Big),
\end{aligned} \tag{C.7}$$

where $\Phi(t) = \int_0^t \phi(s) ds$ is the primitive of the kernel ϕ. This expression can be further simplified using the fact that the normalizing constant on the Gibbs sampler uses only information on the current sampling topic.

Thanks to the analytic form of $\mathbb{P}(z^{s,i}|z^{-(s,i)}, w, \alpha, \eta)$ given by Eqn. (C.4) and by Eqn. (C.5), we have derived a sampling probability for $z^{s,i}$ that depends not only on the topic parameters but also on the Hawkes process itself, which makes it more data-driven.

After achieving the stationary regime for z one can compute the estimators for θ and β as

$$\hat{\theta}_k^t = \frac{n(t,k) + \alpha_k}{\sum_{k'} \left(n(t,k') + \alpha_{k'}\right)} \text{ and } \hat{\beta}_{kj} = \frac{n(j,k) + \eta_j}{\sum_{j'} \left(n(j',k) + \eta_{j'}\right)},$$

where $n(t,k)$ is the number of times a word in message t is of topic k and $n(j,k)$ is the number of times the j-word of the vocabulary was associated with topic k (see [77, 144]).

C.3.1.2 Author-topic model

As already mentioned, since we only have one author per message, the sampling part reduces to sampling only from $z^{s,i}$ instead of sampling from the topic-author couple $z^{s,i}, x^{s,i}$, as in the original topic-author model [265]; nevertheless, we describe here the full form of the collapsed Gibbs sampling for the ATM for the sake of completeness. Thus, we have from [265] that the sampling probabilities for $z^{i,s}$ can be calculated analytically using $z^{-(s,i)}$, $x^{-(s,i)}$, w, a, α and η as

$$\mathbb{P}(z^{s,i} = k, x^{(s,i)} = p | z^{-(s,i)}, x^{-(s,i)}, w, a, \alpha, \eta) \propto \frac{n_{w^i,k}^{-w^i,WT} + \eta_{w^i}}{n_{\cdot,k}^{-w^i,WT} + \sum_j \eta_j} \cdot \frac{n_{p,k}^{-w^i,TA} + \alpha_k}{n_{p,\cdot}^{-w^i,TA} + \sum_{k'} \alpha_{k'}}, \tag{C.8}$$

where

- $n_{w^i,k}^{-w^i,WT}$ is the number of instances of word w^i assigned to topic k, in exception of word w^i in message s,
- $n_{\cdot,k}^{-w^i,WT}$ is the total number of words, in exception of word w^i in message s, that are assigned to topic k,
- $n_{p,k}^{-w^i,TA}$ is the number of words of author p assigned to topic k, in exception of word w^i in message s,
- $n_{p,\cdot}^{-w^i,TA}$ is the total number of words of author p, in exception of word w^i in message s.

Since $\sharp a_s = 1$ for every message s, we have that

$$\mathbb{P}(z^{s,i}, x^{s,i} | z^{-(s,i)}, x^{-(s,i)}, X, w, a, \alpha, \eta) = \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, X, w, a, \alpha, \eta)$$

hence, again by Bayes rule,

$$\begin{aligned}
&\mathbb{P}(z^{s,i}, x^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, a, X, \alpha, \eta) = \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, a, X, \alpha, \eta) \\
&\propto \mathbb{P}(X | z^{s,i}, z^{-(s,i)}, x^{-(s,i)}, w, a, \alpha, \eta) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, a, \alpha, \eta) \\
&= \mathbb{P}(X | z^{s,i}, z^{-(s,i)}) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, a, \alpha, \eta) \\
&= \mathbb{P}(X|Z) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, a, \alpha, \eta) \\
&= L(X|Z) \times \mathbb{P}(z^{s,i} | z^{-(s,i)}, x^{-(s,i)}, w, a, \alpha, \eta),
\end{aligned} \tag{C.9}$$

where the conditional likelihood $L(X|Z)$ is given by Eqn. (C.6), and simplified as in Eqn. (C.7).

Again, thanks to the analytic form of $\mathbb{P}(z^{s,i}|z^{-(s,i)}, x^{-(s,i)}, w, a, \alpha, \eta)$ given by Eqn. (C.8) and by Eqn. (C.9), we have derived a sampling probability for $z^{s,i}$ that depends not only on the topic parameters but also on the Hawkes process itself, which makes it more data-driven.

After achieving the stationary regime for z one can compute the estimators for θ and β as

$$\hat{\theta}_{a,k} = \frac{n^{TA}(a,k) + \alpha_k}{\sum_{k'} n^{TA}(a,k') + \alpha_{k'}} \text{ and}$$
$$\hat{\beta}_{k,j} = \frac{n^{WT}(k,j) + \eta_j}{\sum_{j'} n^{WT}(k,j') + \eta_{j'}},$$

where $n^{TA}(a,k)$ is the number of times a word of author a is of topic k and $n^{WT}(k,j)$ is the number of times the j-word of the vocabulary was associated with topic k (see [265]).

C.3.2 Modified variational Bayes estimation

As stated before, one standard alternative to the Gibbs sampling is the so-called variational approach, where one substitutes the random sampling part with an optimization over some free variables (see [34, 151, 145] for the approach in LDA).

Even though the Gibbs sampling is faster than the optimization approach of the variational Bayes technique, one may still prefer using optimization methods instead of the random sampling. Some of the reasons are:

- The Gibbs sampling needs a burn-in period, where one ignores the first samples, which does not possess any reasonable theoretical guarantee on the minimal number of samples to be ignored.
- After the burn-in period, one considers only every n^{th} sample when averaging values to compute an expectation, since consecutive samples are correlated and form a Markov chain.
- It is difficult to assert that the underlying Markov chain related to the Gibbs sampling procedure indeed converged (the convergence is measured with the autocorrelation function of the samples with different time lags).
- The optimization is easy to perform, and easy to assert convergence.
- The optimization technique uses the entropy function to derive an inferior bound for the function to be maximized, which is easier to replicated to different models than the sampling procedure.

Again, we perform a modified variational Bayes estimation for the topic models in two separate parts. The reason is that although the principles are still the same, the implementations are slightly different. This is due to the introduction of authors on the ATM. Moreover, we also present the calculations for the standard variational Bayes estimation procedure in the ATM, for the sake of completeness. The standard variational Bayes approach for the LDA can be found in [34].

The basic idea behind the standard variational Bayes approach is to derive a lower bound on the log-likelihood of the topic models, by applying Jensen's inequality and introducing the entropy function. For that goal, one introduces free variational parameters following the same distributions as the latent variables, and "breaks" the dependence between the latent variables, generating a simpler Bayesian graphical model.

Then, one estimates the closest variational family to the posterior distribution, with respect to the Kullback-Leibler divergence, and retrieve the free parameters. In this estimation step, one has analytic formulas for the free variational parameters, which must be recalculated in a cyclic way until convergence.

At last, one uses the free parameters to sample the topic model latent variables. For the hyperparameters α and η, one can proceed as in [34, 228] to find a Newton-Raphson algorithm to find the optimal values.

C.3.2.1 Latent Dirichlet Allocation

In order to apply variational Bayes estimation techniques for the LDA, one must first introduce free variational

- Dirichlet variables $\gamma^s = (\gamma_1^s, \cdots, \gamma_K^s)$, $\gamma_k^s \geq 0$ for the message-topic latent variables θ^s,
- discrete variables $\psi^{l,i}$ for the word-topic latent variables $z^{l,i}$, where $\sum_k \psi_k^{l,i} = 1$ and $\psi_k^{l,i} \geq 0$.

One can thus retrieve the random variables θ^s, $z^{l,i}$ as

$$\theta^s \sim Dirichlet(\gamma^s) \quad \text{and} \quad z^{l,i} \sim Discrete(\psi^{l,i}).$$

Secondly, following [34, 145], one must find the minimum distance (given by the Kullback-Liebler divergence) between a variational distribution q and the true posterior $\mathbb{P}(\theta, z|X, w, \alpha, \beta)$, as

$$(\gamma^*, \psi^*) = argmin_{(\gamma,\psi)} d_{KL}(q(\theta, z|\gamma, \psi)|\mathbb{P}(z, \theta|X, w, \alpha, \beta)),$$

where for each message s and each word w^i of message s,

$$q_s(\theta^s, z^s|\gamma^s, \psi^s) = q_s(\theta^s|\gamma^s) \prod_i q_s(z^{s,i}|\psi^{s,i}),$$

with q_s the variational distribution related to message s, which remains[4] in the same exponential family as z and θ, respectively. As one can imagine, the variational distribution q is the proxy of the posterior $\mathbb{P}(\theta, z, X, w|\alpha, \beta)$ in the same exponential family of θ and z.

Our approach, again, makes use of the Hawkes process X to modify the true posterior and introduces a dependence between the dynamics of X and the LDA topic model. To include the Hawkes process X into our posterior, we use Bayes rule as

$$\begin{aligned} \mathbb{P}(\theta, z, w, X|\alpha, \beta) &= \mathbb{P}(X|\theta, z, w, \alpha, \beta).\mathbb{P}(\theta, z, w|\alpha, \beta) \\ &= \mathbb{P}(X|Z).\mathbb{P}(\theta, z, w|\alpha, \beta) = L(X|Z).\mathbb{P}(\theta, z, w|\alpha, \beta), \end{aligned} \tag{C.10}$$

where $L(X|Z)$ is the conditional likelihood of X given Z, as in Eqn. (C.6).

Applying the same methods as in appendix $A.3$ in [34], we have

$$\log \mathbb{P}(X, w|\alpha, \beta) = L(\gamma, \psi; \alpha, \beta) + d_{KL}(q(\theta, z|\gamma, \psi)|\mathbb{P}(\theta, z|X, w, \alpha, \beta)),$$

where

$$\begin{aligned} L(\gamma, \psi; \alpha, \beta) &= \mathbb{E}_q[\log \mathbb{P}(\theta, z, w, X|\alpha, \beta)] - \mathbb{E}_q[q(\theta, z)] \\ &= \mathbb{E}_q[\log L(X|Z)] + \mathbb{E}_q[\log \mathbb{P}(\theta, z, w|\alpha, \beta)] - \mathbb{E}_q[q(\theta, z)] \end{aligned}$$

4. We clearly have $\mathbb{E}_q[z^{s,i}] = \psi^{s,i}$ and $\mathbb{E}_q[\theta^l] = \gamma^l$ (see [34]).

by Eqn. (C.10) and, by Eqn. (C.6),

$$\mathbb{E}_q[\log L(X|Z)] = \mathbb{E}_q[\sum_{n=1}^{X(\tau)} \log(\lambda_{t_n}^{i_n})] - \mathbb{E}_q[\sum_i \int_0^\tau \lambda_t^i dt]. \tag{C.11}$$

One cannot, however, compute analytically Eqn. (C.11). The alternative is to derive a lower bound using the concavity of the logarithm function: let j_l be the user that broadcasted message l. Due to the concavity of the logarithm and the fact that

$$\mathbb{E}_q[Z^{t_l}] = \frac{1}{N_l}\sum_i \mathbb{E}_q[z^{l,i}] = \frac{1}{N_l}\sum_i \psi^{l,i} = \tilde{\psi}^l,$$

one can introduce nonnegative branching variables u_i^t such that $u_{i,0}^t + \sum_{t_l<t,c} u_{i,c,l}^t \tilde{\psi}_c^l = 1$ and bound $\mathbb{E}_q[\log(\lambda_t^i)]$ as

$$\begin{aligned}
\mathbb{E}_q[\log(\lambda_t^i)] &= \mathbb{E}_q[\log(\mu^i + \sum_{j,c,k} J_{i,j} B_{c,k} b_{i,k} \int_0^{t-} \phi(t-s) Z_c^s dX_s^j)] \\
&= \mathbb{E}_q[\log(\mu^i + \sum_{t_l<t,c,k} J_{i,j_l} B_{c,k} b_{i,k} Z_c^{t_l} \phi(t-t_l))] \\
&\geq \sum_{t_l<t,c} u_{i,c,l}^t \mathbb{E}_q[Z_c^{t_l}] \log(J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)) \\
&+ u_{i,0}^t \log(\mu^i) - u_{i,0}^t \log(u_{i,0}^t) - \sum_{t_l<t,c} u_{i,c,l}^t \mathbb{E}_q[Z_c^{t_l}] \log(u_{i,c,l}^t) \\
&= u_{i,0}^t \log(\mu^i) + \sum_{t_l<t,c} u_{i,c,l}^t \tilde{\psi}_c^l \log(J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)) \\
&- u_{i,0}^t \log(u_{i,0}^t) - \sum_{t_l<t,c} u_{i,c,l}^t \tilde{\psi}_c^l \log(u_{i,c,l}^t).
\end{aligned} \tag{C.12}$$

We can find the u_i^t that makes this bound the tightest possible by maximizing it under the constraint $u_{i,0}^t + \sum_{t_l<t,c} u_{i,c,l}^t \tilde{\psi}_c^l = 1$, which gives us

$$\begin{aligned}
u_{i,0}^t &= \frac{\mu^i}{\mu^i + \sum_{t_l<t,c} \tilde{\psi}_c^l J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)} \quad \text{and} \\
u_{i,c,l}^t &= \frac{J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)}{\mu^i + \sum_{t_l<t,c} \tilde{\psi}_c^l J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)}.
\end{aligned}$$

We also trivially have

$$\begin{aligned}
\mathbb{E}_q[\sum_i \int_0^\tau \lambda_t^i dt] &= \sum_i \int_0^\tau \big(\mu^i + \sum_{t_l<t} J_{i,j_l} \sum_{c,k} B_{c,k} b_{i,k} \tilde{\psi}_c^l \phi(t-t_l)\big) dt \\
&= \tau \sum_i \mu^i + \sum_i \sum_{t_l<\tau} \Phi(\tau - t_l) J_{i,j_l} \sum_{c,k} B_{c,k} b_{i,k} \tilde{\psi}_c^l,
\end{aligned} \tag{C.13}$$

where $\Phi(t) = \int_0^t \phi(s) ds$ is the primitive of $\phi(t)$.

Plugging Eqns. (C.12) and (C.13) into $L(\gamma, \psi; \alpha, \beta)$ and deriving with respect to $\psi_c^{l,w}$ we find

$$0 = \partial_{\psi_c^{l,w}} L(\gamma, \psi; \alpha, \beta) = \varphi_c^{l,w} + LDA_c^{l,w},$$

where

$$\varphi_c^{l,w} = \frac{1}{N_l}\Big(\sum_{t_n > t_l} u^n_{i_n,c,l} \log(\frac{J_{i_n,j_l} \sum_k B_{c,k} b_{i_n,k} \phi(t_n - t_l)}{u^n_{i_n,c,l}}) - \Phi(\tau - t_l) \sum_i J_{i,j_l} \sum_k B_{c,k} b_{i,k} \Big),$$

with i_n the user that broadcasted message n, and

$$LDA_c^{l,w} = \Psi(\gamma_c^l) - \Psi(\sum_{c'} \gamma_{c'}^l) + \log \beta_{c,v_w} - \log \psi_c^{l,w} + LM,$$

where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function (see [1] for a numerical implementation using Taylor approximation), $v_w \in \{1, 2, \cdots, W\}$ is the unique index such that $w_v = 1$ and LM is a positive Lagrange multiplier for the constraint $\sum_c \psi_c^{l,w} = 1$ (this result is found in appendix A.3.1 of [34] (above Eqn. (16)).

It is then straightfoward to get

$$\psi_k^{l,w} \propto \beta_{k,v_w} \exp(\varphi_k^{l,w} + \Psi(\gamma_k^l) - \Psi(\sum_{k'} \gamma_{k'}^l)).$$

Since $L(X|Z)$ does not depend on θ (and by consequence on γ) nor β, we have that the updates for γ and β are the same as from the variational Bayes on LDA given by [34], i.e.,

$$\gamma_k^l = \alpha_k + \sum_w \psi_k^{l,w}$$

and (when we consider β fixed instead of following a Dirichlet distribution)

$$\beta_{kj} \propto \sum_{s=1}^{X(\tau)} \sum_{i=1}^{N_s} \psi_k^{s,i} w_j^{s,i}.$$

If we consider $\beta_k \sim Dirichlet(\eta)$ and use a variational parameter ρ_k for each β_k, we get (see [34])

$$\rho_{kj} = \eta_j + \sum_{s=1}^{X(\tau)} \sum_{i=1}^{N_s} \psi_k^{s,i} w_j^{s,i}.$$

C.3.2.2 Author-topic model

We derive here a modified variational Bayes estimation for the author-topic model. We proceed in two steps: first step is to calculate the standard variational Bayes estimation for the author-topic model, without taking into consideration the Hawkes process X; this is done for the sake of completeness, without being able to find it in the literature. Secondly, we introduce the Hawkes process X to derive modified variational Bayes estimates, as previously done in subsubsection C.3.2.1.

Lemma C.1. *Let us define the free variational*

- *Dirichlet variables* $\gamma^a = (\gamma_1^a, \cdots, \gamma_K^a)$, $\gamma_k^a \geq 0$, *for the author-topic latent variables* θ^a,

- *discrete variables* $\psi^{s,i}$ *for the word-topic latent variables* $z^{s,i}$*, where* $\sum_k \psi_k^{s,i} = 1$ *and* $\psi_k^{l,i} \geq 0$*, and*

- *Dirichlet variables* $\rho_k = (\rho_{k,1}, \cdots, \rho_{k,W})$*,* $\rho_{k,j} \geq 0$*, for the topic distributions* β_k*.*

If $\sharp a^t = 1$*, i.e., each message has only one author, the standard variational Bayes free variables for the author-topic model are given by*

$$\gamma_k^a = \alpha_k + \frac{\sum_{s \in A_a} \sum_{w=1}^{N_s} \psi_j^{s,w}}{\sharp A_a}$$

$$\psi_k^{s,w} \propto \exp\left(\Psi(\rho_{k,v_w}) - \Psi(\sum_{j'} \rho_{k,j'}) + \Psi(\gamma_k^{a^s}) - \Psi'(\sum_{k'} \gamma_{k'}^{a^s})\right)$$

$$\rho_{k,j} = \eta_j + \sum_{s=1}^{N_s} \sum_{i=1} \psi_k^{s,i} w_j^{s,i},$$

where a^s *is the author of message* s *and* $A_a = \{s \mid a^s = a\}$ *is the set of messages that have author* $a \in V$ *and* v_w *is the unique index* j *for word* w *such that word* $w_j = 1$*.*

Thus, we can retrieve the random variables θ^a*,* $z^{s,i}$ *and* β_k *as*

$$\theta^a \sim Dirichlet(\gamma^a), \quad z^{s,i} \sim Discrete(\psi^{s,i}) \quad and \quad \beta_k \sim Dirichlet(\rho_k).$$

Proof. We derive a variational Bayes estimation for the author-topic model similar to the one used in [34] for the latent Dirichlet allocation topic model, which is a particular case of [145].

Following appendix $A.3$ of [34], we define the full variational distribution $q(\theta, z, \beta | \gamma, \psi, \rho)$, which is factorized in

$$\begin{aligned} q(\theta, z, \beta | \gamma, \psi, \rho) &= \prod_k q_k(\beta_k | \rho_k) \prod_n q_n(\theta^n | \gamma^n) \prod_{s,i} q_{s,i}(z^{s,i} | \psi^{s,i}) \\ &= q(\beta | \rho) q(\theta | \gamma) q(z | \psi), \end{aligned}$$

where a^s is the author of message s, $w^{s,i}$ are words in message s, and q are the variational distributions[5]. As one can imagine, the variational distribution q is the proxy of the posterior $\mathbb{P}(\theta, z, \beta, w, a | \alpha, \eta)$ in the same exponential family of θ, z and β.

In order to use the variational approach in the appendix $A.3$ of [34], we need to find the minimum distance (given by the Kullback-Liebler divergence) between the variational distribution q and the true posterior $\mathbb{P}(\theta, z, \beta | w, a, \alpha, \eta)$.

Let $w^s = (w^{s,1}, \cdots, w^{s,N_s})$ be the words in message s, with respective topic latent variables $z^s = (z^{s,1}, \cdots, z^{s,N_s})$, and let us define the variational free energy

$$L(\gamma, \psi, \rho; \alpha, \eta) = \mathbb{E}_q[\log \mathbb{P}(w, a, z, \theta, \beta | \alpha, \eta)] - \mathbb{E}_q[\log q(\theta, z, \beta)]. \tag{C.14}$$

5. One can see that since q remains in the same exponential family as z and θ, we have $\mathbb{E}_q[z^{s,i}] = \psi^{s,i}$, $\mathbb{E}_q[\theta^a] = \gamma^a$ and $\mathbb{E}_q[\beta_k] = \rho_k$ (see [34]).

Applying the same methods as in appendix $A.3$ in [34], we have that

$$\begin{aligned}
\log \mathbb{P}(w, a|\alpha, \eta) &= \sum_s \log \mathbb{P}(w^s, a^s|\alpha, \eta) = \sum_s \log \int \sum_{z^s} \mathbb{P}(w^s, a^s, z^s, \theta, \beta|\alpha, \eta) d\theta d\beta \\
&= \sum_s \log \int \sum_{z^s} \frac{\mathbb{P}(w^s, a^s, z^s, \theta|\alpha, \eta) q(\theta) q(z^s) q(\beta) d\theta d\beta}{q(\theta) q(z^s) q(\beta)} \\
&\geq \sum_s \mathbb{E}_q[\log \mathbb{P}(w^s, a^s, z^s, \theta, \beta|\alpha, \eta)] - \mathbb{E}_q[\log \big(q(\theta) q(z^s) q(\beta)\big)] \\
&= \mathbb{E}_q[\log \mathbb{P}(w, a, z, \theta, \beta|\alpha, \eta)] - \mathbb{E}_q[\log q(\theta, z, \beta)] \\
&= L(\gamma, \psi, \rho; \alpha, \eta)
\end{aligned}$$

by Jensen's inequality for the logarithm function and by the fact that the messages are independent given the authors.

Since by Bayes rule

$$\log \mathbb{P}(w, a|\alpha, \eta) = \log \mathbb{P}(w, a, \theta, z, \beta|\alpha, \eta) - \log \mathbb{P}(\theta, z, \beta|w, a, \alpha, \eta),$$

we have that

$$\begin{aligned}
\log \mathbb{P}(w, a|\alpha, \eta) &= \mathbb{E}_q[\log \mathbb{P}(w, a|\alpha, \eta)] \\
&= \mathbb{E}_q[\log \mathbb{P}(w, a, \theta, z, \beta|\alpha, \eta)] - \mathbb{E}_q[\log \mathbb{P}(\theta, z, \beta|w, a, \alpha, \eta)],
\end{aligned}$$

thus thanks to Eqn. (C.14)

$$\begin{aligned}
\log \mathbb{P}(w, a|\alpha, \eta) &= \mathbb{E}_q[\log \mathbb{P}(w, a, \theta, z, \beta|\alpha, \eta)] - \mathbb{E}_q[\log \mathbb{P}(\theta, z, \beta|w, a, \alpha, \eta)] \\
&= L(\gamma, \psi, \rho; \alpha, \eta) + \mathbb{E}_q[\log q(\theta, z, \beta)] - \mathbb{E}_q[\log \mathbb{P}(\theta, z, \beta|w, a, \alpha, \eta)] \\
&= L(\gamma, \psi, \rho; \alpha, \eta) + d_{KL}(q(\theta, z, \beta|\gamma, \psi, \rho)|\mathbb{P}(\theta, z, \beta|w, a, \alpha, \eta)),
\end{aligned}$$

where d_{KL} is the Kullback-Leibler divergence between probability distributions. This implies that in order to minimize the Kullback-Leibler divergence between the posterior $\mathbb{P}(\theta, z, \beta|w, a, \alpha, \eta)$ and the variational distribution $q(\theta, z, \beta|\gamma, \psi, \rho)$, one may simply maximize the free energy $L(\gamma, \psi, \rho; \alpha, \eta)$.

On the other hand, following the graphical model in Figure C.2, we have that (we used the fact that $\sharp a^s = 1$)

$$\begin{aligned}
\mathbb{P}(w, a, \theta, z, \beta|\alpha, \eta) &= \mathbb{P}(\theta|\alpha)\mathbb{P}(z|a, \theta)\mathbb{P}(w|z, \beta)\mathbb{P}(\beta|\eta) \\
&= \prod_{n \in V} \mathbb{P}(\theta^n|\alpha) \prod_s \prod_i \mathbb{P}(z^{s,i}|\theta^{a^s}) \mathbb{P}(w^{s,i}|z^{s,i}, \beta) \prod_k \mathbb{P}(\beta_k|\eta).
\end{aligned}$$

Now, we calculate each term separately, following appendix $A.3$ of [34]:

Since θ^n are Dirichlet random variables and the variational distributions q remain in the same family, we have

$$\begin{aligned}
\mathbb{E}_{q_n}[\log \mathbb{P}(\theta^n|\alpha)] &= \log \Gamma(\sum_k \alpha_k) - \sum_k \log \Gamma(\alpha_k) + \sum_k (\alpha_k - 1) \mathbb{E}_{q_n}[\log \theta_k^n] \\
&= \log \Gamma(\sum_k \alpha_k) - \sum_k \log \Gamma(\alpha_k) + \sum_k (\alpha_k - 1) \Big(\Psi(\gamma_k^n) - \Psi(\sum_{k'} \gamma_{k'}^n) \Big)
\end{aligned}$$

and

$$\mathbb{E}_{q_n}[\log q_n(\theta^n)] = \log\Gamma(\sum_k \gamma_k^n) - \sum_k \log\Gamma(\gamma_k^n) + \sum_k (\gamma_k^n - 1)\Big(\Psi(\gamma_k^n) - \Psi(\sum_{k'} \gamma_{k'}^n)\Big),$$

where $\Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ is the digamma function (see [1] for a numerical implementation using Taylor approximation).

Since z are discrete random variables, we have by Eqn. (C.2) that

$$\mathbb{E}_{q_{s,i}}[\log \mathbb{P}(z^{s,i}|\theta^{a^s})] = \sum_k \psi_k^{s,i} \mathbb{E}_{q_{s,i}}[\log \theta_k^{a^s}] = \sum_k \psi_k^{s,i} \Big(\Psi(\gamma_k^{a^s}) - \Psi(\sum_{k'} \gamma_{k'}^{a^s})\Big)$$

and

$$\mathbb{E}_{q_{s,i}}[\log q_{s,i}(z^{s,i})] = \sum_k \psi_k^{s,i} \log \psi_k^{s,i},$$

Since β_k are also Dirichlet random variables, we have as in θ^i

$$\begin{aligned}\mathbb{E}_{q_k}[\log \mathbb{P}(\beta_k|\eta)] &= \log\Gamma(\sum_j \eta_j) - \sum_j \log\Gamma(\eta_j) \\ &\quad + \sum_j (\eta_j - 1)\Big(\Psi(\rho_{k,j}) - \Psi(\sum_{j'} \rho_{k,j'})\Big)\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}_{q_k}[\log q_k(\beta_k)] &= \log\Gamma(\sum_j \rho_{k,j}) - \sum_j \log\Gamma(\rho_{k,j}) \\ &\quad + \sum_j (\rho_{k,j} - 1)\Big(\Psi(\rho_{k,j}) - \Psi(\sum_{j'} \rho_{k,j'})\Big),\end{aligned}$$

Finally, we have by Eqn. (C.3) that

$$\begin{aligned}\mathbb{E}_q[\log \mathbb{P}(w^{s,i}|z^{s,i},\beta)] &= \sum_j \sum_k w_j^{s,i} \psi_k^{s,i} \mathbb{E}_{q_k}[\log \beta_{k,j}] \\ &= \sum_j \sum_k w_j^{s,i} \psi_k^{s,i} \Big(\Psi(\rho_{k,j}) - \Psi(\sum_{j'} \rho_{k,j'})\Big).\end{aligned}$$

Putting all these terms together, we get that

$$
\begin{aligned}
&L(\gamma,\psi,\rho;\alpha,\eta) = \mathbb{E}_q[\log \mathbb{P}(\theta,z,w,a,\beta|\alpha,\eta)] - \mathbb{E}_q[\log q(\theta,z,\beta)] \\
&= \sum_n \Bigg[\log\Gamma(\sum_k \alpha_k) - \sum_k \log\Gamma(\alpha_k) + \sum_k (\alpha_k - 1)\Big(\Psi(\gamma_k^n) - \Psi(\sum_{k'} \gamma_{k'}^n)\Big) \\
&- \log\Gamma(\sum_k \gamma_k^n) + \sum_k \log\Gamma(\gamma_k^n) - \sum_k (\gamma_k^n - 1)\Big(\Psi(\gamma_k^n) - \Psi(\sum_{k'} \gamma_{k'}^n)\Big)\Bigg] \\
&+ \sum_k \Bigg[\log\Gamma(\sum_j \eta_j) - \sum_j \log\Gamma(\eta_j) + \sum_j (\eta_j - 1)\Big(\Psi(\rho_{k,j}) - \Psi(\sum_{j'} \rho_{k,j'})\Big) \\
&- \log\Gamma(\sum_j \rho_{k,j}) + \sum_j \log\Gamma(\rho_{k,j}) - \sum_j (\rho_{k,j} - 1)\Big(\Psi(\rho_{k,j}) - \Psi(\sum_{j'} \rho_{k,j'})\Big)\Bigg] \\
&+ \sum_{s,i} \Bigg[\sum_k \psi_k^{s,i}\Big(\Psi(\gamma_k^{a^s}) - \Psi(\sum_{k'} \gamma_{k'}^{a^s})\Big) - \sum_k \psi_k^{s,i} \log \psi_k^{s,i} \\
&+ \sum_j \sum_k w_j^{s,i} \psi_k^{s,i}\Big(\Psi(\rho_{k,j}) - \Psi(\sum_{j'} \rho_{k,j'})\Big)\Bigg].
\end{aligned}
$$

We have then

$$
\begin{aligned}
0 &= \partial_{\gamma_k^a} L(\gamma,\psi,\rho;\alpha,\beta) \\
&= \sum_{s\in A_a} \Bigg(\Psi'(\gamma_k^a)[\alpha_k + \sum_{w=1}^{N_s} \psi_k^{s,w} - \gamma_k^a] - \Psi'(\sum_{k'} \gamma_{k'}^a) \sum_j [\alpha_j + \sum_{w=1}^{N_s} \psi_j^{s,w} - \gamma_j^a] \Bigg),
\end{aligned}
$$

where $A_a = \{s \mid a^s = a\}$, which gives us the following updates for γ_k^a

$$
\gamma_k^a = \alpha_k + \frac{\sum_{s\in A_a} \sum_{w=1}^{N_s} \psi_j^{s,w}}{\sharp A_a}.
$$

For $\psi^{s,w}$, we have that

$$
\begin{aligned}
\partial_{\psi_k^{s,w}} L(\gamma,\psi,\rho;\alpha,\beta) &= \Psi(\gamma_k^{a_s}) - \Psi'(\sum_{k'} \gamma_{k'}^{a_s}) \\
&+ \Psi(\rho_{k,v_w}) - \Psi(\sum_{j'} \rho_{k,j'}) - \log(\psi_k^{s,w}) + LM^{s,w} = 0,
\end{aligned} \tag{C.15}
$$

where v_w is the index for word w in the dictionary, a_s is the author of message s and $LM^{s,w}$ is a Lagrange multiplier for the constraint $\sum_k \psi_k^{s,w} = 1$. We have thus the update for $\psi_k^{s,i}$

$$
\psi_k^{s,w} \propto \exp\Big(\Psi(\rho_{k,v_w}) - \Psi(\sum_{j'} \rho_{k,j'}) + \Psi(\gamma_k^{a^s}) - \Psi'(\sum_{k'} \gamma_{k'}^{a^s})\Big).
$$

And for $\rho_{k,j}$, we have that

$$
\partial_{\rho_{k,j}} L(\gamma,\psi,\rho;\alpha,\beta) = \Big(\Psi'(\rho_{k,j}) - \Psi'(\sum_{j'} \rho_{k,j'})\Big)[\eta_j + \sum_s \sum_{i=1}^{N_s} \psi_k^{s,i} w_j^{s,i} - \rho_{k,j}] = 0,
$$

which gives us the following update for $\rho_{k,j}$

$$\rho_{k,j} = \eta_j + \sum_s \sum_{i=1}^{N_s} \psi_k^{s,i} w_j^{s,i}.$$

□

Following subsubsection C.3.2.1, we introduce the free variational

- Dirichlet variables $\gamma^a = (\gamma_1^a, \cdots, \gamma_K^a)$, $\gamma_k^a \geq 0$, for the author-topic latent variables θ^a, and
- discrete variables $\psi^{l,i}$ for the word-topic latent variables $z^{l,i}$, where $\sum_k \psi_k^{l,i} = 1$ and $\psi_k^{l,i} \geq 0$

We can thus retrieve the random variables θ^a, $z^{l,i}$ as

$$\theta^a \sim Dirichlet(\gamma^a) \quad \text{and} \quad z^{l,i} \sim Discrete(\psi^{l,i}).$$

Our approach, again, makes use of the Hawkes process X to modify the true posterior and introduces a dependence between the dynamics of X and the author-topic model. To include the Hawkes process X into our posterior, we use Bayes rule as

$$\begin{aligned} &\mathbb{P}(\theta, z, w, a, X|\alpha, \beta) = \mathbb{P}(X|\theta, z, w, a, \alpha, \beta).\mathbb{P}(\theta, z, w, a|\alpha, \beta) \\ &= \mathbb{P}(X|Z).\mathbb{P}(\theta, z, w, a|\alpha, \beta) = L(X|Z).\mathbb{P}(\theta, z, w, a|\alpha, \beta), \end{aligned} \tag{C.16}$$

where $L(X|Z)$ is the conditional likelihood of X given Z, as in Eqn. (C.6).

Applying the same methods as in appendix $A.3$ in [34] and subsubsection C.3.2.1, we have

$$\log \mathbb{P}(X, w, a|\alpha, \beta) = L(\gamma, \psi; \alpha, \beta) + d_{KL}(q(\theta, z|\gamma, \psi)|\mathbb{P}(\theta, z|X, w, a, \alpha, \beta)),$$

where

$$\begin{aligned} L(\gamma, \psi; \alpha, \beta) &= \mathbb{E}_q[\log \mathbb{P}(\theta, z, w, a, X|\alpha, \beta)] - \mathbb{E}_q[q(\theta, z)] \\ &= \mathbb{E}_q[\log L(X|Z)] + \mathbb{E}_q[\log \mathbb{P}(\theta, z, w, a|\alpha, \beta)] - \mathbb{E}_q[q(\theta, z)] \end{aligned}$$

by Eqn. (C.16) and, by Eqn. (C.6),

$$\mathbb{E}_q[\log L(X|Z)] = \mathbb{E}_q[\sum_{n=1}^{X(\tau)} \log(\lambda_{t_n}^{i_n})] - \mathbb{E}_q[\sum_i \int_0^\tau \lambda_t^i dt]. \tag{C.17}$$

Since one cannot compute analytically Eqn. (C.17), we derive a lower bound using the concavity of the logarithm function: let j_l be the user that broadcasted message l. Due to the concavity of the logarithm and the fact that

$$\mathbb{E}_q[Z^{t_l}] = \frac{1}{N_l} \sum_i \mathbb{E}_q[z^{l,i}] = \frac{1}{N_l} \sum_i \psi^{l,i} = \tilde{\psi}^l,$$

one can introduce nonnegative branching variables u_i^t such that $u_{i,0}^t + \sum_{t_l<t,c} u_{i,c,l}^t \tilde{\psi}_c^l = 1$ and bound $\mathbb{E}_q[\log(\lambda_t^i)]$ in the following way:

$$\begin{aligned}
\mathbb{E}_q[\log(\lambda_t^i)] &= \mathbb{E}_q[\log(\mu^i + \sum_{t_l<t} \sum_{c,k} J_{i,j_l} B_{c,k} b_{i,k} \phi(t-t_l) Z_c^{t_l})] \\
&\geq \sum_{t_l<t,c} u_{i,c,l}^t \mathbb{E}_q[Z_c^{t_l}] \log(J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)) \\
&+ u_{i,0}^t \log(\mu^i) - u_{i,0}^t \log(u_{i,0}^t) - \sum_{t_l<t,c} u_{i,c,l}^t \mathbb{E}_q[Z_c^{t_l}] \log(u_{i,c,l}^t) \\
&= u_{i,0}^t \log(\mu^i) + \sum_{t_l<t,c} u_{i,c,l}^t \tilde{\psi}_c^l \log(J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)) \\
&- u_{i,0}^t \log(u_{i,0}^t) - \sum_{t_l<t,c} u_{i,c,l}^t \tilde{\psi}_c^l \log(u_{i,c,l}^t). \qquad \text{(C.18)}
\end{aligned}$$

We can find the u_i^t that makes this bound the tightest possible by maximizing it under the constraint $u_{i,0}^t + \sum_{t_l<t,c} u_{i,c,l}^t \tilde{\psi}_c^l = 1$, which gives us

$$u_{i,0}^t = \frac{\mu^i}{\mu^i + \sum_{t_l<t,c} \tilde{\psi}_c^l J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)}$$

and

$$u_{i,c,l}^t = \frac{J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)}{\mu^i + \sum_{t_l<t,c} \tilde{\psi}_c^l J_{i,j_l} \sum_k B_{c,k} b_{i,k} \phi(t-t_l)}.$$

We also trivially have

$$\mathbb{E}_q[\sum_i \int_0^\tau \lambda_t^i dt] = \tau \sum_i \mu^i + \sum_i \sum_{t_l<\tau} \Phi(\tau - t_l) J_{i,j_l} \sum_{c,k} B_{c,k} b_{i,k} \tilde{\psi}_c^l,$$

where $\Phi(t) = \int_0^t \phi(s) ds$ is the primitive of $\phi(t)$.

Plugging Eqns. (C.18) and (C.19) into $L(\gamma, \psi; \alpha, \beta)$ and deriving with respect to $\psi_c^{l,w}$ we find

$$0 = \partial_{\psi_c^{l,w}} L(\gamma, \psi; \alpha, \beta) = \varphi_c^{l,w} + AT_c^{l,w},$$

where

$$\varphi_c^{l,w} = \frac{1}{N_l} \Big(\sum_{t_s>t_l} u_{i_s,c,l}^s \log(\frac{J_{i_s,j_l} \sum_k B_{c,k} b_{i_s,k} \phi(t_s - t_l)}{u_{i_s,c,l}^s}) - \sum_i \Phi(\tau - t_l) J_{i,j_l} \sum_k B_{c,k} b_{i,k} \Big),$$

with i_n the user that broadcasted message n,

$$AT_c^{l,w} = \log \beta_{c,v_w} + \Psi(\gamma_c^{i_l}) - \Psi'(\sum_{c'} \gamma_{c'}^{i_l}) - \log(\psi_c^{l,w}) + LM^{l,w},$$

as in Eqn. (C.15) of lemma C.1, $v_w \in \{1, 2, \cdots, W\}$ is the unique index such that $w_v = 1$ and $LM^{l,w}$ is a positive Lagrange multiplier for the constraint $\sum_c \psi_c^{l,w} = 1$.

It is then straightforward to get

$$\psi_k^{l,w} \propto \beta_{k,v_w} \exp(\varphi_k^{l,w} + \Psi(\gamma_k^{i_l}) - \Psi(\sum_{k'} \gamma_{k'}^{i_l})).$$

Since $L(X|Z)$ does not depend on θ (and by consequence on γ) nor β, we have that the updates for γ and β are the same as in lemma C.1.

Again, if we consider $\beta_k \sim Dirichlet(\eta)$ and use a variational parameter ρ_k for each β_k, we get by lemma C.1

$$\rho_{k,j} = \eta_j + \sum_s \sum_{i=1}^{N_s} \psi_k^{s,i} w_j^{s,i}.$$

C.4 Additional remarks

In this thesis we used the latent Dirichlet allocation and the author-topic topic models to take into consideration the "randomness" of the topics in each message broadcasted by users in social networks. The latent Dirichlet allocation topic model, developed by Blei *et al.* in [34], is one of the most used and understood topic models, and its usage can be verified in several domains outside of text mining, such as collaborative filtering, image retrieval, bioinformatics, etc., where the same can be said about the author-topic model.

However, LDA and ATM are rather simple models, albeit their common use. One of their strengths - and the choice of LDA and ATM as the topic models in this thesis - stems exactly from this simplicity, which allows one to extend and complexify both topic models to one's desire. For example, some of these extensions are:

- hierarchical topic models [32], where topics are joined together in a hierarchy by using the nested Chinese restaurant process,
- LDA-dual model [271], where one uses a corpus in which a document includes two types of information (e.g., words and names),
- Hidden Markov Model LDA [129], where one distinguishes between different types of words (e.g., function words and content words),
- supervised topic models [31] and semi-supervised topic models [219], where documents can possess observed labels,
- dynamic topic models [30, 297], where topics can evolve in time,
- extensions of LDA with Hierarchical Dirichlet process mixture model [281], which allows the number of topics to be unbounded and learned from data [280],
- Spatial LDA [299], where for instance one automatically puts natural images into categories, such as "bedroom" or "forest", by treating an image as a document, and small patches of the image as words [199].

APPENDIX D

Tools used in chapter 4

We discuss in this appendix the two necessary tools in order to develop our trend detection algorithm of chapter 4: how to rescale nearly unstable Hawkes processes, as in [163], and how to detect the maximum of a mean-reverting scalar Itô diffusion, as in [97].

D.1 Rescaling Hawkes process

D.1.1 Introduction

According to chapter 4, we have a multivariate linear Hawkes process $X_t^{i,k}$ with intensity of the form

$$\lambda_t^{i,k} = \mu^{i,k} + \sum_c \sum_j B_{c,k} J_{i,j} \int_0^{t-} \phi(t-s) dX_s^{j,c}, \tag{D.1}$$

which in matrix form can be seen as

$$\lambda_t = \mu + J(\phi * dX)_t B,$$

where μ is the intrinsic rate of dissemination, J is the user-user interaction matrix, B is the topic-topic interaction matrix and $(\phi * dX)_t$ is the $N \times K$ convolution matrix defined as $(\phi * dX)_t^{i,k} = \int_0^t \phi(t-s) dX_s^{i,k}$.

Hawkes processes have two distinct regimes: stable and unstable - their boundaries are given by lemma 24. However, the estimation of the Hawkes parameters J and B in finance led practitioners to believe that most of Hawkes processes are stable, but near the unstability regime, i.e., $sp(J)sp(B)||\phi||_1 \sim 1$, as for example in [136].

Indeed, one of the most well documented facts in high frequency finance is the long memory of markets, see for example [206]. Since Hawkes processes are the point process equivalent of autoregressive processes, they exhibit short range dependence, failing to reproduce this classical long memory empirical feature. Moreover, since the estimation of the Hawkes parameters is done under a finite time horizon, a condition of the form $sp(J)sp(B)||\phi||_1 \sim 1$ may imply that the estimation is done during an insufficiently small time horizon, hence counting too many jumps per timeframe and "pushing" the Hawkes parameters near the instability regime.

One solution to remedy to this is by rescaling the Hawkes process so as to have less jumps per timeframe. Jaisson and Rosenbaum study in [163] the rescaling of a nearly unstable one-dimensional Hawkes process X_t^τ, $t \in [0, \tau]$, and show that, under mild assumptions, the rescaling of X_t^τ converges to a Cox-Ingersoll-Ross (CIR) process [69]. CIR processes are used in finance to

model interest rates. They have two interesting properties: a CIR process is nonnegative (even strictly positive under certain conditions) and mean-reverting, i.e., it is an ergodic process which admits a long-term average, and its evolution is a fluctuation around this well-defined average.

D.1.2 Assumptions

In order to prove our main rescaling convergence result, we make the following assumptions:

Assumption D.1. *The temporal kernel $\phi(t)$ is an exponential function with timescale parameter ω_τ*

$$\phi(t) = e^{-\omega_\tau t}.\mathbb{I}_{\{t>0\}}.$$

Remark: Assumption D.1 is in fact a simplifying one, and one may use any temporal kernel satisfying the hypothesis in [163].

Assumption D.2. *The interaction matrices J and B can be diagonalized into $J = v^{-1}\nu v$ and $B = \rho D \rho^{-1}$ and $B^T \otimes J$ has only one maximal eigenvalue. Thus, in light of the decomposition for J and B, we have that J has left-eigenvectors the rows of v, denoted by v_i^T, with associated eigenvalues ν_i; and B has right-eigenvectors the columns of ρ, denoted by ρ_k, with associated eigenvalues $D_{k,k}$, i.e., v^T is the $N \times N$ matrix and ρ is the $K \times K$ matrix*

$$v^T = \begin{pmatrix} | & \dots & | \\ v_1^T & \dots & v_N^T \\ | & \dots & | \end{pmatrix} \quad \text{and} \quad \rho = \begin{pmatrix} | & \dots & | \\ \rho_1 & \dots & \rho_K \\ | & \dots & | \end{pmatrix}.$$

Since the eigenvalues of $B^T \otimes J$ are of the form $v_i D_{k,k}$, (i,k), we assume without loss of generality that $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_N$ and $D_{1,1} > D_{2,2} \geq D_{3,3} \geq \cdots \geq D_{K,K}$, and that the largest eigenvalues of J and B satisfy $\nu_1 > 0$ and $D_{1,1} > 0$.

Moreover, we also have that v_1^T and ρ_1 have nonnegative entries by the Perron-Frobenius theorem, since J and B have nonnegative entries (this result remains true for the leading right-eigenvector of J and the leading left-eigenvector of B as well).

Remark: The assumption that J and B can be diagonalized is in fact a simplifying one. One could use the Jordan blocks of J and B, on the condition that there exists only one maximal eigenvalue for $B^T \otimes J$. This assumption is verified if, for example, the graph associated with B is strongly connected; which means that every topic influences the other topics, even if it is in an undirected fashion (by influencing topics that will, in their turn, influence other topics, and so on). One can also develop a theory in the case of multiple maximal eigenvalues, but it will be much more complicated and the associated stochastic control problem has not yet been solved analytically, hence numerical methods must be employed.

Assumption D.3. *We have that the timescale parameter ω_τ satisfies, for some $\lambda > 0$,*

$$\tau(1 - \frac{\nu_1 D_{1,1}}{\omega_\tau}) \to \lambda$$

when $\tau \to \infty$, which implies

$$\omega_\tau \searrow \nu_1 D_{1,1}.$$

D.1.3 Rescaling theorem

The multidimensional case presented here works in a similar fashion, which is achieved by the following theorem:

Theorem D.1. *Let X be the multivariate Hawkes process in $[0,\tau]$ with intensity given by Eqn. (D.1), and let $\varphi_t^{i,k} = v_i^T \frac{\lambda_{\tau t}}{\tau} \rho_k$, where v_i^T and ρ_k are defined in assumption D.2.*

Under assumptions D.1, D.2 and D.3 we have that

- *If $(i,k) \neq (1,1)$ then $\varphi_t^{i,k}$ converges in law to 0 for the Skorokhod topology in $[0,1]$ when $\tau \to \infty$.*
- *Let v_1^T and $\tilde{v}_1$ be the leading left and right eigenvectors of J associated with the eigenvalue $\nu_1 > 0$, let ρ_1 and $\tilde{\rho}_1^T$ be the leading right and left eigenvectors of B associated with the eigenvalue $D_{1,1} > 0$, and define $\pi = (\sum_i v_{1,i}^2 \tilde{v}_{i,1})(\sum_k \rho_{k,1}^2 \tilde{\rho}_{1,k})$.*

  Thus, $\varphi_t^{1,1}$ converges in law to the CIR process C_t for the Skorokhod topology in $[0,1]$ when $\tau \to \infty$, where C_t satisfies the following stochastic differential equation

$$\begin{cases} dC_t &= \lambda \nu_1 D_{1,1} (\frac{\mu}{\lambda} - C_t) dt + \nu_1 D_{1,1} \sqrt{\pi} \sqrt{C_t} dW_t, \ t \in [0,1] \\ C_0 &= 0, \end{cases}$$

  where W_t is a standard Brownian motion.

D.1.4 Proof of theorem D.1

We now proceed to the proof of theorem D.1, following the ideas in [162]. We provide here a sketch of the proof:

1. We start by writing the equations satisfied by the rescaled intensities $\varphi_t^{i,k} = v_i^T \frac{\lambda_{\tau t}}{\tau} \rho_k$ and study their first-order properties.
2. Secondly, we define the new martingales $B_t^{i,k}$ and show that they converge to a standard Brownian motion.
3. Thirdly, we rewrite $\varphi_t^{1,1}$ in a more suitable form, with remainder terms U_t and V_t, and we show that they converge to 0.
4. Finally, we apply the convergence theorem 5.4 of [184] for limits of stochastic integrals with semimartingales.

D.1.5 Rescaling the Hawkes intensity

Let us begin by defining the one-dimensional stochastic processes

$$\tilde{\lambda}_t^{i,k} = v_i^T \lambda_t \rho_k,$$

which satisfy the one-dimensional equations

$$\begin{aligned} \tilde{\lambda}_t^{i,k} &= v_i^T \mu \rho_k + v_i^T J (\phi * dX)_t B \rho_k \\ &= \tilde{\mu}^{i,k} + \nu_i D_{k,k} (\phi * \tilde{\lambda}^{i,k})_t + \nu_i D_{k,k} (\phi * v_i^T dM \rho_k)_t, \end{aligned} \tag{D.2}$$

with $M_t = X_t - \int_0^t \lambda_s ds$ the compensated martingale associated with the Hawkes process X and $\tilde{\mu}^{i,k} = v_i^T \mu \rho_k$. Using lemma 2.1 of [163], we have that

$$\tilde{\lambda}_t^{i,k} = \tilde{\mu}^{i,k} + \tilde{\mu}^{i,k} \int_0^t \Psi_{i,k}(t-s)ds + \int_0^t \Psi_{i,k}(t-s) v_i^T dM_s \rho_k, \tag{D.3}$$

where

$$\Psi_{i,k}(t) = \sum_{n \geq 1} (\nu_i D_{k,k} \phi(t))^{*n},$$

with the n^{th} convolution operator defined as

$$(\nu_i D_{k,k}\phi(t))^{*1} = \nu_i D_{k,k}\phi(t), \quad \text{and} \quad (\nu_i D_{k,k}\phi(t))^{*n} = ((\nu_i D_{k,k}\phi)^{*(n-1)} * \nu_i D_{k,k}\phi)_t.$$

We have the following lemma for the convolutions $\Psi_{i,k}$:

Lemma D.1. *Let $\Psi_{i,k}(t) = \sum_{n\geq 1}(\nu_i D_{k,k}\phi(t))^{*n}$, then under assumption D.1 we have that*

$$\Psi_{i,k}(t) = \nu_i D_{k,k} e^{-\omega_\tau (1 - \frac{\nu_i D_{k,k}}{\omega_\tau})t}.$$

Moreover, under assumptions D.2 and D.3 we have that

$$\Psi_{1,1}(\tau t) \to \nu_1 D_{1,1} e^{-\nu_1 D_{1,1} \lambda t}$$

uniformly in $[0,1]$ when $\tau \to \infty$, and that there exists a constant $L > 0$ such that for $(i,k) \neq (1,1)$ we have

$$\int_0^t \Psi_{i,k}(\tau(t-s))ds \leq \frac{L}{\tau}. \tag{D.4}$$

Proof. Under assumption D.1, we have that

$$(\nu_i D_{k,k}\phi(t))^{*2} = (\nu_i D_{k,k})^2 \int_0^t e^{-\omega_\tau (t-s)} e^{-\omega_\tau s} ds = (\nu_i D_{k,k})^2 t e^{-\omega_\tau t}$$
$$\Rightarrow (\nu_i D_{k,k}\phi(t))^{*n} = (\nu_i D_{k,k})^n \frac{t^{n-1}}{(n-1)!} e^{-\omega_\tau t},$$

hence

$$\Psi_{i,k}(t) = e^{-\omega_\tau t} \sum_{n \geq 1} \nu_i^n D_{k,k}^n \frac{t^{n-1}}{(n-1)!} = \nu_i D_{k,k} e^{-(1 - \frac{\nu_i D_{k,k}}{\omega_\tau})\omega_\tau t}.$$

Now, under assumptions D.2 and D.3, we have that $\tau(1 - \frac{\nu_1 D_{1,1}}{\omega_\tau}) \to \lambda$ and $\omega_\tau \to \nu_1 D_{1,1}$, which implies that there exists a constant $\underline{\lambda} > 0$ such that for every $(i,k) \neq (1,1)$

$$\omega_\tau (1 - \frac{\nu_i D_{k,k}}{\omega_\tau}) \geq \underline{\lambda} > 0 \quad \text{and} \quad \tau(1 - \frac{\nu_i D_{k,k}}{\omega_\tau}) \to \infty.$$

Firstly, for $t \in [0,1]$, using the Lipschitz continuity of e^{-t} we have that

$$|\Psi_{1,1}(\tau t) - \nu_1 D_{1,1} e^{-\nu_1 D_{1,1} \lambda t}| \leq (\nu_1 D_{1,1})^2 \lambda (\sup_{s \in [0,1]} e^{-\nu_1 D_{1,1} \lambda s}) t |\omega_\tau \tau (1 - \frac{\nu_1 D_{1,1}}{\omega_\tau}) - \nu_1 D_{1,1} \lambda|$$
$$\leq (\nu_1 D_{1,1})^2 \lambda |\omega_\tau \tau (1 - \frac{\nu_1 D_{1,1}}{\omega_\tau}) - \nu_1 D_{1,1} \lambda| \to 0$$

when $\tau \to \infty$, which implies that $\Psi_{1,1}(\tau t) \to \nu_1 D_{1,1} e^{-\nu_1 D_{1,1} \lambda t}$ uniformly in $[0,1]$.

At last, for $(i,k) \neq (1,1)$, we have that

$$\int_0^t \Psi_{i,k}(\tau(t-s))ds = \nu_i D_{k,k} \frac{1 - e^{-t\tau\omega_\tau(1-\frac{\nu_i D_{k,k}}{\omega_\tau})}}{\tau\omega_\tau(1-\frac{\nu_i D_{k,k}}{\omega_\tau})} \leq \frac{L}{\tau}$$

where $L > 0$ is a large enough positive constant. □

Let us now define the one-dimensional rescaled stochastic processes, for $t \in [0,1]$,

$$\varphi_t^{i,k} = \frac{v_i^T \lambda_{\tau t} \rho_k}{\tau},$$

which clearly satisfies

$$\varphi_t = v \frac{\lambda_{\tau t}}{\tau} \rho. \tag{D.5}$$

We have the following lemma concerning the first order properties of φ_t:

Lemma D.2. *Let us define the $1 \times N$ row vector $v_i^{\odot 2}$ such that $(v_i^{\odot 2})_j = v_{i,j}^2$ and the $K \times 1$ vector $\rho_k^{\odot 2}$ such that $(\rho_k^{\odot 2})_c = \rho_{c,k}^2$. We have*

1. *$\varphi_t^{i,k}$ satisfies the following equation*

$$\varphi_t^{i,k} = \tilde{\mu}^{i,k}(\frac{1}{\tau} + \int_0^t \Psi_{i,k}(\tau(t-s))ds) + \int_0^t \Psi_{i,k}(\tau(t-s)) \sqrt{(v_i^{\odot 2})^T \frac{\lambda_{\tau s}}{\tau} \rho_k^{\odot 2}} dB_s^{i,k}, \tag{D.6}$$

   where

$$B_t^{i,k} = \sqrt{\tau} \int_0^t \frac{v_i^T dM_{\tau s} \rho_k}{\sqrt{(v_i^{\odot 2})^T \lambda_{\tau s} \rho_k^{\odot 2}}} = \int_0^t \frac{v_i^T dM_{\tau s} \rho_k}{\sqrt{(v_i^{\odot 2})^T v^{-1} \varphi_s \rho^{-1} \rho_k^{\odot 2}}}$$

   is a L^2 martingale.

2. *If $(i,k) \neq (1,1)$, then*

$$\mathbb{E}[\varphi_t^{i,k}] \leq \frac{L}{\tau},$$

   where $L > 0$ is a large enough positive constant. Moreover, we also have that

$$\mathbb{E}[\varphi_t^{1,1}] \leq L.$$

Proof. 1. We have that

$$\begin{aligned}
\varphi_t^{i,k} &= \frac{1}{\tau}\tilde{\mu}^{i,k} + \frac{1}{\tau}\tilde{\mu}^{i,k} \int_0^{\tau t} \Psi_{i,k}(\tau t - s)ds + \frac{1}{\tau} \int_0^{\tau t} \Psi_{i,k}(\tau t - s) v_i^T dM_s \rho_k \\
&= \frac{1}{\tau}\tilde{\mu}^{i,k} + \tilde{\mu}^{i,k} \int_0^t \Psi_{i,k}(\tau(t-s))ds + \int_0^t \Psi_{i,k}(\tau(t-s)) v_i^T dM_{\tau s} \rho_k \\
&= \frac{1}{\tau}\tilde{\mu}^{i,k} + \tilde{\mu}^{i,k} \int_0^t \Psi_{i,k}(\tau(t-s))ds \\
&+ \int_0^t \Psi_{i,k}(\tau(t-s)) \sqrt{(v_i^{\odot 2})^T \frac{\lambda_{\tau s}}{\tau}}_s \rho_k^{\odot 2} \frac{v_i^T \sqrt{\tau} dM_{\tau s} \rho_k}{\sqrt{(v_i^{\odot 2})^T \lambda_{\tau s} \rho_k^{\odot 2}}} \\
&= \frac{1}{\tau}\tilde{\mu}^{i,k} + \tilde{\mu}^{i,k} \int_0^t \Psi_{i,k}(\tau(t-s))ds + \int_0^t \Psi_{i,k}(\tau(t-s)) \sqrt{(v_i^{\odot 2})^T \frac{\lambda_{\tau s}}{\tau}}_s \rho_k^{\odot 2} dB_s^{i,k}.
\end{aligned}$$

As $\frac{\lambda_{\tau s}}{\tau} = v^{-1}\varphi_s\rho^{-1}$ by Eqn. (D.5), we have the result.

2. Since $B_t^{i,k}$ is a martingale, we have that

$$\begin{aligned}\mathbb{E}[\varphi_t^{i,k}] &= \frac{1}{\tau}\tilde{\mu}^{i,k} + \frac{1}{\tau}\tilde{\mu}^{i,k}\int_0^{\tau t}\Psi_{i,k}(\tau t - s)ds \\ &= \frac{1}{\tau}\tilde{\mu}^{i,k} + \tilde{\mu}^{i,k}\int_0^{t}\Psi_{i,k}(\tau(t - s))ds,\end{aligned}$$

which together with lemma D.1 gives us the result.

□

Remark: We can assume, without loss of generality, that there exists a $c > 0$ such that

- $v_1^T\mu\rho_1 = \tilde{\mu}^{1,1} \geq c$, since $v_1^T\mu\rho_1 = \tilde{\mu}^{1,1} = 0 \Rightarrow \mathbb{E}[\varphi_t^{1,1}] = 0$ by lemma D.2, which implies $\varphi_t^{1,1} = 0$ almost surely for all $t \geq 0$ by the fact that $\varphi_t^{1,1} \geq 0$ for all $t \geq 0$, and
- $\min_{(i,k)}(v^{\odot 2})_i^T\mu\rho_k^{\odot 2} \geq c$, which implies for all (i,k) that $(v^{\odot 2})_i^T\lambda_s\rho_k^{\odot 2} \geq (v^{\odot 2})_i^T\mu\rho_k^{\odot 2} \geq c > 0$.

D.1.6 Second order properties

We study now the second order properties of $B_t^{i,k}$ and $\varphi_t^{i,k}$, with the help of a classical lemma, which we do not prove.

Lemma D.3. *Let $f : \mathbb{M}_{N\times K}(\mathbb{R}^+) \to \mathbb{R}^+$ and $g : \mathbb{M}_{N\times K}(\mathbb{R}^+) \to \mathbb{R}^+$ be functions satisfying for some constant $C > 0$*

$$|f(\varphi_t)| \leq C(1 + ||\varphi_t||) \quad \textit{and} \quad |g(\varphi_t)| \leq C(1 + ||\varphi_t||),$$

let $h : \mathbb{R} \to \mathbb{R}$ and $r : \mathbb{R} \to \mathbb{R}$ be continuous functions and let Z_t^1 and Z_t^2 be L^2 martingales such that $[Z^1, Z^2]_t = t + M_t$, where M_t is a martingale.
Defining $z_t^1 = \int_0^t h(s)f(\varphi_s)dZ_s^1$ and $z_t^2 = \int_0^t r(s)g(\varphi_s)dZ_s^2$ we have that

$$\mathbb{E}[z_t^1 z_t^2] = \int_0^t h(s)r(s)\mathbb{E}[f(\varphi_s)g(\varphi_s)]ds.$$

Moreover, if Z_t is a L^2 semimartingale, we have that the stochastic process $Y_t = \int_0^t h(s)f(\varphi_s)dZ_s$ satisfies

$$[Y]_t = \int_0^t h^2(s)f^2(\varphi_s)d[Z]_s \quad \textit{and} \quad \mathbb{E}[Y_t^2] \leq \mathbb{E}[[Y]_t].$$

Regarding the second order properties of $B_t^{i,k}$ and $\varphi_t^{i,k}$, we have the following lemma:

Lemma D.4. *For each (i,k) let $[B^{i,k}]_t$ be the quadratic variation of the martingale $B_t^{i,k}$. We have that*

1.

$$[B^{i,k}]_t = t + \frac{1}{\tau}\int_0^{\tau t}\frac{(v_i^{\odot 2})^T dM_s\rho_k^{\odot 2}}{(v_i^{\odot 2})^T\lambda_s\rho_k^{\odot 2}}. \tag{D.7}$$

2.
$$\mathbb{E}[(\varphi_t^{i,k})^2] \leq L \tag{D.8}$$
for a constant $L > 0$.

3. Moreover, if $(i,k) \neq (1,1)$, then
$$\mathbb{E}[(\varphi_t^{i,k})^2] \leq \frac{L}{\tau^2}$$
for a constant $L > 0$.

Proof. 1. Since the Hawkes process X does not have more than one jump at each time, we have that
$$[M^{i,k}, M^{j,c}]_t = X_t^{i,k} \mathbb{I}_{\{(i,k)=(j,c)\}},$$
which by its turn implies
$$[v_i^T M \rho_k]_t = (v_i^{\odot 2})^T X_t \rho_k^{\odot 2}.$$
We have by lemma D.2 that
$$B_t^{i,k} = \sqrt{\tau} \int_0^t \frac{v_i^T dM_{\tau s} \rho_k}{\sqrt{(v_i^{\odot 2})^T \lambda_{\tau s} \rho_k^{\odot 2}}} = \frac{1}{\sqrt{\tau}} \int_0^{\tau t} \frac{v_i^T dM_s \rho_k}{\sqrt{(v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}}},$$
hence
$$\begin{aligned}[B^{i,k}]_t &= \frac{1}{\tau} \int_0^{\tau t} \frac{(v_i^{\odot 2})^T dX_s \rho_k^{\odot 2}}{(v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}} = \frac{1}{\tau} \int_0^{\tau t} \frac{(v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}}{(v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}} ds + \frac{1}{\tau} \int_0^{\tau t} \frac{(v_i^{\odot 2})^T dM_s \rho_k^{\odot 2}}{(v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}} \\ &= t + \frac{1}{\tau} \int_0^{\tau t} \frac{(v_i^{\odot 2})^T dM_s \rho_k^{\odot 2}}{(v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}}.\end{aligned}$$

2. Using the fact that $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, we have by Eqn. (D.6) and lemma D.1 that
$$\begin{aligned}(\varphi_t^{i,k})^2 &\leq 3\Bigg(\frac{(\tilde{\mu}^{i,k})^2}{\tau^2} + (\tilde{\mu}^{i,k})^2 \Big(\int_0^t \Psi_{i,k}(\tau(t-s)) ds \Big)^2 \\ &\quad + \Big(\int_0^t \Psi_{i,k}(\tau(t-s)) \sqrt{(v_i^{\odot 2})^T v^{-1} \varphi_s \rho^{-1} \rho_k^{\odot 2}} dB_s^{i,k} \Big)^2 \Bigg) \\ &\leq L' + 3 \Big(\int_0^t \Psi_{i,k}(\tau(t-s)) \sqrt{(v_i^{\odot 2})^T v^{-1} \varphi_s \rho^{-1} \rho_k^{\odot 2}} dB_s^{i,k} \Big)^2.\end{aligned}$$
Since $\Psi_{i,k}(\tau(t-s)) = \Psi_{i,k}(\tau t) \Psi_{i,k}(-\tau s)$, we have then
$$\begin{aligned}(\varphi_t^{i,k})^2 &\leq L' + 3 \Big(\int_0^t \Psi_{i,k}(\tau(t-s)) \sqrt{(v_i^{\odot 2})^T v^{-1} \varphi_s \rho^{-1} \rho_k^{\odot 2}} dB_s^{i,k} \Big)^2 \\ &= L' + 3 \Psi_{i,k}^2(\tau t) (Z_t^{i,k})^2,\end{aligned}$$
with $Z_t^{i,k} = \int_0^t \Psi_{i,k}(-\tau s) \sqrt{(v_i^{\odot 2})^T v^{-1} \varphi_s \rho^{-1} \rho_k^{\odot 2}} dB_s^{i,k}$ a martingale. By lemma D.3 we have by lemma D.2 that
$$\begin{aligned}\mathbb{E}[(\varphi_t^{i,k})^2] &\leq L' + 3 \Psi_{i,k}^2(\tau t) \int_0^t \Psi_{i,k}^2(-\tau s) (v_i^{\odot 2})^T v^{-1} \mathbb{E}[\varphi_s] \rho^{-1} \rho_k^{\odot 2} ds \\ &= L' + 3 \int_0^t \Psi_{i,k}^2(\tau(t-s)) (v_i^{\odot 2})^T v^{-1} \mathbb{E}[\varphi_s] \rho^{-1} \rho_k^{\odot 2} ds \leq L.\end{aligned}$$

3. For $(i,k) \neq (1,1)$, we have from Eqn. (D.3), lemma D.1 and the inequality $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ that

$$(\tilde{\lambda}_t^{i,k})^2 \leq L' + 3(\int_0^t \Psi_{i,k}(t-s) v_i^T dM_s \rho_k)^2.$$

One can promptly see by lemma D.2 that $\mathbb{E}[\lambda_t^{i,k}] = (v^{-1}\mathbb{E}[\tilde{\lambda}_t]\rho^{-1})_{i,k} \leq L'''$. Hence using lemma D.3 and the same calculation of the previous item gives

$$\begin{aligned}\mathbb{E}[(\tilde{\lambda}_t^{i,k})^2] &\leq L' + 3\int_0^t \Psi_{i,k}^2(t-s)(v_i^{\odot 2})^T \mathbb{E}[\lambda_s]\rho_k^{\odot 2} ds \\ &\leq L''(1 + \int_0^t \Psi_{i,k}^2(t-s)ds) \leq L\end{aligned}$$

for a constant $L > 0$. Thus $\mathbb{E}[(\varphi_t^{i,k})^2] = \frac{\mathbb{E}[(\tilde{\lambda}_{\tau t}^{i,k})^2]}{\tau^2} \leq \frac{L}{\tau^2}$, as desired.

□

We derive next the convergence properties of the martingales $B_t^{i,k}$ and the rescaled process $\varphi_t^{i,k}$, $(i,k) \neq (1,1)$, which use two lemmas. The first one is

Lemma D.5. *Let X be a $N \times K$ matrix with nonnegative entries, $v^T \neq 0$ be a $1 \times N$ row vector with nonnegative entries and $\rho \neq 0$ be a $K \times 1$ vector with nonnegative entries. Then*

$$(v^{\odot 2})^T X \rho^{\odot 2} \leq ||v||.||\rho||.v^T X \rho.$$

Proof. Define the row vector $\tilde{v}^T = \frac{v^T}{||v||}$ and the vector $\tilde{\rho} = \frac{\rho}{||\rho||}$, such that $\tilde{v}_i^T \leq 1$ and $\tilde{\rho}_k \leq 1$, which implies $(\tilde{v}_i)^2 \leq \tilde{v}_i$ and $(\tilde{\rho}_k)^2 \leq \tilde{\rho}_k$. Then

$$\begin{aligned}(v^{\odot 2})^T X \rho^{\odot 2} &= ||v||^2.||\rho||^2 \sum_{i,k} \tilde{v}_i^2 X_{i,k} \tilde{\rho}_k^2 \leq ||v||^2.||\rho||^2. \sum_{i,k} \tilde{v}_i X_{i,k} \tilde{\rho}_k \\ &= ||v||.||\rho||. \sum_{i,k} v_i X_{i,k} \rho_k = ||v||.||\rho|| v^T X \rho.\end{aligned}$$

□

The second one is[1]

Lemma D.6. *Let $f_\tau : \mathbb{R}^+ \to \mathbb{R}$ be a sequence of functions such that*

1. *There exists a constant $C > 0$ such that $\sup_\tau \sup_{t \in \mathbb{R}} |f_\tau(t)| \leq C$,*
2. *There exists a constant $C > 0$ such that for all τ*

$$|f_\tau(t) - f_\tau(s)| \leq C\tau|t-s|,$$

3. *For any $0 < \varepsilon < 1$, there exists $C_\varepsilon > 0$ such that for every t, s*

$$\sup_\tau \int_{\mathbb{R}} (f_\tau(t-u) - f_\tau(s-u))^2 du \leq C_\varepsilon |t-s|^{1-\varepsilon},$$

1. This lemma can be proven using the ideas in [163] (see the proof of the convergence for the rescaled process $(Y_t^T)_{t \in [0,1]}$ at the beginning of page 18, corollaries 4.1, 4.2, 4.3, 4.4 and lemma 4.7).

4. $\int_{\mathbb{R}^+} f_\tau^2(s)ds \to 0$ *when* $\tau \to \infty$, *and*
5. *There exists a constant* $C > 0$ *such that* $\sup_\tau |\hat{f}_\tau(z)| \leq C(|\frac{1}{z}| \wedge 1)$.

Let $g : \mathcal{M}_{N\times K}(\mathbb{R}^+) \to \mathbb{R}$ *be a function satisfying for some constant* $C > 0$

$$|g(\varphi_t)| \leq C(1 + ||\varphi_t||),$$

and define $Y_t^{i,k,\tau} = \int_0^t f_\tau(t-s)g(\varphi_s)dB_s^{i,k}$.

We have that $Y_t^{i,k,\tau}$ *converges in law to* 0 *for the Skorohod topology in* $[0,1]$ *when* $\tau \to \infty$.

Now, for the convergence of $B_t^{i,k}$ and $\varphi_t^{i,k}$:

Lemma D.7. *We have that*

1. *For every* (i,k), $B_t^{i,k}$ *converges in law to a standard Brownian motion for the Skorohod topology in* $[0,1]$ *when* $\tau \to \infty$.
2. *If* $(i,k) \neq (1,1)$, *then* $\varphi_t^{i,k}$ *converges in law to* 0 *for the Skorohod topology in* $[0,1]$ *when* $\tau \to \infty$.

Proof. 1. By Eqn. (D.7) we have for $t \in [0,1]$

$$\begin{aligned}
\mathbb{E}[([B^{i,k}]_t - t)^2] &= \mathbb{E}[(\frac{1}{\tau}\int_0^{\tau t} \frac{(v_i^{\odot 2})^T dM_s \rho_k^{\odot 2}}{(v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}})^2] \leq \frac{1}{\tau^2}\mathbb{E}[\int_0^{\tau t} \frac{d[(v_i^{\odot 2})^T M \rho_k^{\odot 2}]_s}{\left((v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}\right)^2}] \\
&= \frac{1}{\tau^2}\mathbb{E}[\int_0^{\tau t} \frac{(v_i^{\odot 4})^T \lambda_s \rho_k^{\odot 4} ds}{\left((v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}\right)^2}] \\
&\leq ||v_i^{\odot 2}||.||\rho_k^{\odot 2}||.\frac{1}{\tau^2}\mathbb{E}[\int_0^{\tau t} \frac{ds}{(v_i^{\odot 2})^T \lambda_s \rho_k^{\odot 2}}] \\
&\leq L\frac{1}{\tau^2}\int_0^{\tau t} dt \leq \frac{L}{\tau}
\end{aligned}$$

for some $L > 0$ by lemma D.5.

Thus, by Markov's inequality we have that for all $\varepsilon > 0$ and for all $t \in [0,1]$

$$\mathbb{P}(|[B^{i,k}]_t - t| \geq \varepsilon) \leq \frac{L}{\tau\varepsilon^2} \to 0 \text{ when } \tau \to \infty,$$

which shows that, for every $t \in [0,1]$, $[B^{i,k}]_t$ converges in probability towards t when $\tau \to \infty$. Since $B^{i,k}$ has uniformly bounded jumps because X and λ have uniformly bounded jumps, we have by theorem $VIII.3.11$ of [161] that $B_t^{i,k}$ converges in law to a standard Brownian motion for the Skorohod topology in $[0,1]$ when $\tau \to \infty$.

2. Since $\sup_{t\in[0,1]} \mathbb{E}[\varphi_t^{i,k}] \to 0$ when $\tau \to \infty$ by lemma D.2, we have by Eqn. (D.6) that we only need to prove the convergence of $Z_t^{i,k} = \int_0^t \Psi_{i,k}(\tau(t-s))g(\varphi_s)dB_s^{i,k}$, where $g(\varphi_s) = \sqrt{(v_i^{\odot 2})^T v^{-1}\varphi_s \rho^{-1}\rho_k^{\odot 2}}$ satisfies $|g(\varphi_s)| \leq C(1 + ||\varphi_s||)$ for some $C > 0$.

Since $\Psi_{i,k}(\tau(t-s))$ is an exponential function by lemma D.1, we have that assumption D.3 implies that $\Psi_{i,k}(\tau(t-s))$ satisfies all hypothesis of lemma D.6, and as consequence we have that $Z_t^{i,k}$ converges in law to 0 for the Skorohod topology in $[0,1]$ when $\tau \to \infty$, which concludes the proof.

□

D.1.7 Convergence of $\varphi_t^{1,1}$

After studying the asymptotic behavior of the martingale B_t and the rescaled processes $\varphi_t^{i,k}$ for $(i,k) \neq (1,1)$, we study the asymptotic behavior of $\varphi_t^{1,1}$. We start by rewriting it in a more convenient form, using Eqn. (D.6):

$$\begin{aligned}\varphi_t^{1,1} = \tilde{\mu}^{1,1}(\frac{1}{\tau} + \int_0^t \Psi_{1,1}(\tau(t-s))ds) + \int_0^t \nu_1 D_{1,1} e^{-\nu_1 D_{1,1}\lambda(t-s)}\sqrt{\pi\varphi_s^{1,1}}dB_s^{1,1} \\ + U_t + V_t,\end{aligned} \tag{D.9}$$

where $\pi = (\sum_i v_{1,i}^2 (v^{-1})_{i,1})(\sum_k \rho_{k,1}^2 (\rho^{-1})_{1,k})$,

$$U_t = \int_0^t \Psi_{1,1}(\tau(t-s))(\sqrt{(v_1^{\odot 2})^T v^{-1}\varphi_s \rho^{-1}\rho_1^{\odot 2}} - \sqrt{\pi\varphi_s^{1,1}})dB_s^{1,1}. \tag{D.10}$$

and

$$V_t = \int_0^t \left(\Psi_{1,1}(\tau(t-s)) - \nu_1 D_{1,1} e^{-\nu_1 D_{1,1}\lambda(t-s)}\right)\sqrt{\pi\varphi_s^{1,1}}dB_s^{1,1} \tag{D.11}$$

We begin by studying the asymptotic behavior of U_t in Eqn. (D.10) and V_t in (D.11), which need an additional lemma:

Lemma D.8. *Let $\pi = (\sum_i v_{1,i}^2 (v^{-1})_{i,1})(\sum_k \rho_{k,1}^2 (\rho^{-1})_{1,k})$. We have that*

$$(v_1^{\odot 2})^T v^{-1}\varphi_t \rho^{-1}\rho_1^{\odot 2} \leq \pi\varphi_t^{1,1} + L\sum_{(i,k)\neq(1,1)} \varphi_t^{i,k}$$

for some constant $L > 0$.

Proof. Let us define the $1 \times N$ row vector $V^T = (v_1^{\odot 2})^T v^{-1}$ and the $K \times 1$ vector $R = \rho^{-1}\rho_1^{\odot 2}$, such that

$$V_j^T = \sum_i v_{1,i}^2 v_{i,j}^{-1} \quad \text{and} \quad R_c = \sum_k \rho_{c,k}^{-1}\rho_{k,1}^2.$$

Thus

$$\begin{aligned}(v_1^{\odot 2})^T v^{-1}\varphi_t \rho^{-1}\rho_1^{\odot 2} = \sum_{j,c} V_j^T \varphi_t^{j,c} R_c = \pi\varphi_t^{1,1} + \sum_{(j,c)\neq(1,1)} V_j^T \varphi_t^{j,c} R_c \\ \leq \pi\varphi_t^{1,1} + L\sum_{(i,k)\neq(1,1)} \varphi_t^{i,k}\end{aligned}$$

for a constant $L > 0$. □

Lemma D.9. *We have that U_t defined in Eqn. (D.10) converges in law to 0 for the Skorohod topology in $[0,1]$ when $\tau \to \infty$.*

Proof. Let us define the martingale $Z_t = \int_0^t \Psi_{1,1}(-\tau s)(\sqrt{(v_1^{\odot 2})^T v^{-1}\varphi_s \rho^{-1}\rho_1^{\odot 2}} - \sqrt{\pi\varphi_s^{1,1}})dB_s^{1,1}$, such that $U_t = \Psi_{1,1}(\tau t)Z_t$. Using the product formula for semimartingales and the fact that $\Psi_{1,1}$ has bounded variation, we have that

$$U_t = \int_0^t \partial_t \Psi_{1,1}(\tau s)Z_s ds + \int_0^t \Psi_{1,1}(\tau s)dZ_s = V_t + W_t,$$

where $V_t = \int_0^t \partial_t \Psi_{1,1}(\tau s) Z_s ds$ has bounded variation and $W_t = \int_0^t \Psi_{1,1}(\tau s) dZ_s$ is a martingale with quadratic variation $[W]_t$ satisfying by lemma D.3

$$\begin{aligned}
[W]_t &= \int_0^t \Psi_{1,1}^2(\tau s) d[Z]_s \\
&= \int_0^t \Psi_{1,1}^2(\tau s)\Psi_{1,1}^2(-\tau s)(\sqrt{(v_1^{\odot 2})^T v^{-1}\varphi_s \rho^{-1}\rho_1^{\odot 2}} - \sqrt{\pi\varphi_s^{1,1}})^2 d[B^{1,1}]_s \\
&= \int_0^t (\sqrt{(v_1^{\odot 2})^T v^{-1}\varphi_s \rho^{-1}\rho_1^{\odot 2}} - \sqrt{\pi\varphi_s^{1,1}})^2 d[B^{1,1}]_s.
\end{aligned}$$

Thus, using the fact that $\sqrt{a+b} - \sqrt{b} \leq \frac{a}{2\sqrt{b}}$ for $a, b > 0$, we have by lemma D.4

$$\begin{aligned}
\mathbb{E}[[W]_t] &= \mathbb{E}[\int_0^t (\sqrt{(v_1^{\odot 2})^T v^{-1}\varphi_s \rho^{-1}\rho_1^{\odot 2}} - \sqrt{\pi\varphi_s^{1,1}})^2 ds] \\
&\leq \mathbb{E}[\int_0^t \frac{((v_1^{\odot 2})^T v^{-1}\varphi_s \rho^{-1}\rho_1^{\odot 2} - \pi\varphi_s^{1,1})^2}{4\pi\varphi_s^{1,1}} ds] \\
&\leq L^2 \mathbb{E}[\int_0^t \frac{(\sum_{(i,k)\neq(1,1)} \varphi_s^{i,k})^2}{4\pi\varphi_s^{1,1}} ds]
\end{aligned}$$

for some constant $L > 0$ by lemma D.8. Since $\varphi_s^{1,1} \geq \frac{\tilde{\mu}^{1,1}}{\tau} \geq \frac{c}{\tau} > 0$, we have by lemma D.4 that for $t \in [0,1]$

$$\begin{aligned}
\mathbb{E}[[W]_t] &\leq \frac{\tau L^2}{4\pi c} \mathbb{E}[\int_0^t (\sum_{(i,k)\neq(1,1)} \varphi_s^{i,k})^2 ds] \leq \frac{\tau L(NK-1)}{4\pi c} \mathbb{E}[\int_0^t \sum_{(i,k)\neq(1,1)} (\varphi_s^{i,k})^2 ds] \\
&= \frac{\tau L(NK-1)}{4\pi c} \int_0^t \sum_{(i,k)\neq(1,1)} \mathbb{E}[(\varphi_s^{i,k})^2] ds \\
&\leq \frac{\tau L^2 (NK-1)^2}{4\pi c} \int_0^t \frac{1}{\tau^2} ds = \frac{L' t}{\tau} \leq \frac{L'}{\tau}
\end{aligned}$$

for $L' = \frac{L^2(NK-1)^2}{4\pi c}$.

Thus, by Markov's inequality we have that for all $\varepsilon > 0$ and for all $t \in [0,1]$

$$\mathbb{P}([W]_t \geq \varepsilon) \leq \frac{L'}{\varepsilon}(\frac{1}{\tau^2} + \frac{1}{\tau}) \to 0 \quad \text{when} \quad \tau \to \infty,$$

which proves that $[W]_t$ converges in probability to 0 for all $t \geq 0$.

Since W has uniformly bounded jumps, we have by theorem $VIII$.3.11 of [161] that W_t converges in law to 0 for the Skorohod topology in $[0,1]$ when $\tau \to \infty$.

Now, regarding V_t, we have that since $|\partial_t \Psi_{1,1}(\tau t)| \leq C$ for some constant $C > 0$,

$$\mathbb{E}[(V_t - V_s)^2] = \mathbb{E}[(\int_s^t \partial_t \Psi_{1,1}(\tau u) Z_u du)^2] \leq C^2 (t-s)^2 \mathbb{E}[\sup_{u\in[s,t]} Z_u^2] \leq C'(t-s)^2 \mathbb{E}[[Z]_t]$$

by the Burkholder-Davis-Gundy inequality. Since by lemma D.3

$$[Z]_t = \int_0^t \Psi_{1,1}^2(-\tau s)(\sqrt{(v_1^{\odot 2})^T v^{-1}\varphi_s \rho^{-1}\rho_1^{\odot 2}} - \sqrt{\pi\varphi_s^{1,1}})^2 d[B^{1,1}]_s$$

and $\Psi_{1,1}(-\tau s) \leq C$ for some constant $C > 0$, we have using the same calculations as before and choosing $s = 0$ that for $t \in [0, 1]$

$$\mathbb{E}[V_t^2] \leq C' t^2 \mathbb{E}[[Z]_t] \leq \frac{C''}{\tau^2},$$

which easily implies that $(V_{t_1}, \cdots, V_{t_n}) \to 0$ in distribution for every $(t_1, \cdots, t_n) \in [0, 1]^n$ when $\tau \to \infty$, i.e., we have the convergence of the finite-dimensional distribution of V_t to 0 when $\tau \to \infty$.

Moreover, since $\mathbb{E}[(V_t - V_s)^2] \leq C'''(t - s)^2$, we have by the Kolmogorov criterion for tightness that V_t is tight for the Skorohod topology in $[0, 1]$, which implies that V_t converges in law to 0 for the Skorohod topology in $[0, 1]$ when $\tau \to \infty$.

Hence, we clearly have that $U_t = V_t + W_t$ converges in law to 0 for the Skorohod topology in $[0, 1]$ when $\tau \to \infty$. □

Lemma D.10. *We have that V_t defined in Eqn. (D.11) converges in law to 0 for the Skorohod topology in $[0, 1]$ when $\tau \to \infty$.*

Proof. Define the function

$$f_\tau(t) = \Psi_{1,1}(\tau t) - \nu_1 D_{1,1} e^{-\nu_1 D_{1,1} \lambda t} = \nu_1 D_{1,1} \left(e^{-\omega_\tau \tau (1 - \frac{\nu_1 D_{1,1}}{\omega_\tau}) t} - e^{-\nu_1 D_{1,1} \lambda t} \right).$$

By assumption D.3, that there exists a $C > 0$ such that

1. $\sup_\tau \sup_t |f_\tau(t)| \leq C$,
2. Since f_τ is a difference of exponential functions, we can assume without loss of generality that $|\hat{f}_\tau(z)| \leq C(|\frac{1}{z}| \wedge 1)$,
3. Applying lemma 4.7 of [163] we have that for any $0 < \varepsilon < 1$, there exists $C_\varepsilon > 0$ such that for every t, s
$$\sup_\tau \int_{\mathbb{R}} (f_\tau(t - u) - f_\tau(s - u))^2 du \leq C_\varepsilon |t - s|^{1-\varepsilon},$$
4. Since
$$f_\tau^2(t) = \nu_1^2 D_{1,1}^2 \left(\Psi_{1,1}^2(\tau t) + e^{-2\nu_1 D_{1,1} \lambda t} - 2\Psi_{1,1}(\tau t) e^{-\nu_1 D_{1,1} \lambda t} \right),$$
we have that
$$\int_{\mathbb{R}^+} f_\tau^2(t) dt = \nu_1^2 D_{1,1}^2 \left(\frac{1}{2\omega_\tau \tau (1 - \frac{\nu_1 D_{1,1}}{\omega_\tau})} + \frac{1}{2\nu_1 D_{1,1} \lambda} - 2 \frac{1}{\omega_\tau \tau (1 - \frac{\nu_1 D_{1,1}}{\omega_\tau}) + \nu_1 D_{1,1} \lambda} \right) \to 0.$$
5. Since, for $\alpha > 0$, $e^{-\alpha t}$ satisfies $|e^{-\alpha t} - e^{-\alpha s}| \leq \alpha |t - s|$, we easily have that there exists a constant $C > 0$ such that
$$|f_\tau(t) - f_\tau(s)| \leq C\tau |t - s|.$$

Hence, f_τ satisfies all hypothesis of lemma D.6. Moreover, $g(\varphi_s) = \sqrt{\pi\varphi_s^{1,1}}$ satisfies

$$|g(\varphi_t)| \leq C(1 + ||\varphi_t||).$$

We can thus apply lemma D.6 to conclude the proof. □

We have arrived to the final step of the proof: by lemma D.1, we have that $\tilde{\mu}^{1,1}\int_0^t \Psi_{1,1}(\tau(t-s))ds$ converges uniformly in $[0,1]$ to $\tilde{\mu}^{1,1}\int_0^t \nu_1 D_{1,1} e^{-\nu_1 D_{1,1}\lambda(t-s)}ds = \tilde{\mu}^{1,1}(\frac{1-e^{-\nu_1 D_{1,1}\lambda t}}{\lambda})$, when $\tau \to \infty$.

Moreover, by lemma D.7 we have that $B_t^{1,1}$ converges in law to a standard Brownian motion for the Skorohod topology in $[0,1]$ when $\tau \to \infty$, and by lemmas D.9 and D.10 we have that U_t and V_t converge in law to 0 for the Skorohod topology in $[0,1]$ when $\tau \to \infty$.

As in [163], since U_t and V_t converge to a deterministic limit, we get the convergence in law, for the product topology, of the triple $(U_t, V_t, B_t^{1,1})$ to $(0,0,W_t)$ with W a standard Brownian motion. The components of $(0,0,W_t)$ being continuous, the last convergence also takes place for the Skorohod topology on the product space.

Thus, we have by theorem 5.4 of [184] that $\varphi_t^{1,1}$ converges in law to the limit process C_t for the Skorohod topology in $[0,1]$ when $\tau \to \infty$, where C_t is the unique solution of

$$C_t = \tilde{\mu}^{1,1}(\frac{1-e^{-\nu_1 D_{1,1}\lambda t}}{\lambda}) + \nu_1 D_{1,1}\int_0^t e^{-\nu_1 D_{1,1}\lambda(t-s)}\sqrt{\pi C_s}dW_s,$$

where W_t is a standard Brownian motion.

By a simple calculation, we have that C_t satisfies the following stochastic differential equation

$$\begin{aligned}
dC_t &= \nu_1 D_{1,1}\tilde{\mu}^{1,1}e^{-\nu_1 D_{1,1}\lambda t}dt + \nu_1 D_{1,1}\left(-\nu_1 D_{1,1}\lambda\int_0^t e^{-\nu_1 D_{1,1}\lambda(t-s)}\sqrt{\pi C_s}dW_s dt + \sqrt{\pi C_t}dW_t\right)\\
&= \nu_1 D_{1,1}\tilde{\mu}^{1,1}e^{-\nu_1 D_{1,1}\lambda t}dt + \nu_1 D_{1,1}\left((\tilde{\mu}^{1,1}(1-e^{-\nu_1 D_{1,1}\lambda t}) - \lambda C_t)dt + \sqrt{\pi C_t}dW_t\right)\\
&= \nu_1 D_{1,1}\lambda(\frac{\tilde{\mu}^{1,1}}{\lambda} - C_t)dt + \nu_1 D_{1,1}\sqrt{\pi C_t}dW_t.
\end{aligned}$$

Remark: One promptly has that the columns of v^{-1} are the right-eigenvectors of J and that the rows of ρ^{-1} are the left-eigenvectors of B, thus $\pi > 0$ can be rewritten as

$$\pi = (\sum_i v_{1,i}^2 \tilde{v}_{i,1})(\sum_k \rho_{k,1}^2 \tilde{\rho}_{1,k}),$$

where v_1^T is the leading left-eigenvector of J, $\tilde{v}_1$ is the right-eigenvector of J, ρ_1 is the leading right-eigenvector of B and $\tilde{\rho}_1^T$ is the leading left-eigenvector of B. Moreover, by the Perron-Frobenius theorem we have that v, $\tilde{v}$, ρ and $\tilde{\rho}$ have nonnegative entries.

Remark: In the one-dimensional case, we clearly have that $\pi = 1$, retrieving thus the same result as in [163].

D.2 Detecting the maximum of mean-reverting Itô diffusions

Another fundamental tool to our trend detection algorithm is the detection of maximum in mean-reverting Itô diffusions, as proposed in [94, 97].

A (time-homogeneous) Itô diffusion X_t is a stochastic process in $\mathbb{R}^n$, solution of a stochastic differential equation (SDE) of the form

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,\ X_0 \in \mathbb{R}^n,$$

where W_t is a standard n-dimensional Brownian motion. The term μ is called the drift coefficient, and σ is called the diffusion coefficient. The drift coefficient is responsible for the "deterministic" infinitesimal change in $X_{t+dt} - X_t$, of size $\mu(X_t)dt$, and σ is responsible for the randomness of this change, having variance $\sigma^2(X_t)dt$ and zero mean.

Under some regularity conditions (Lipschitz continuity of μ and σ), the Itô diffusion has almost surely continuous paths, admits a unique strong solution and is strongly Markovian (see [244]). Itô diffusions are very similar to the Langevin equation in physics, and have applications in a multitude of domains, such as quantitative finance, control theory, neuroscience, harmonic analysis, telecommunications, etc...

Here, we enunciate briefly the work of Espinosa and Touzi developed in [97], where they adopt a stochastic control framework to detect the maximum of a nonnegative scalar Itô diffusion: They consider the following Itô diffusion

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t,\ X_0 > 0, \tag{D.12}$$

with[2] $\mu(x) \leq 0$, W_t a standard Brownian motion, and the following optimal stopping problem:

$$V_0 = \inf_{\theta \in \mathcal{T}_0} \mathbb{E}[l(X^*_{T_0} - X_\theta)], \tag{D.13}$$

where

- $l(x) \geq 0$ is a convex cost function,
- $X^*_t = \sup_{s \leq t} X_s$ is the running maximum of X_t,
- $T_y = \inf\{t > 0 \mid X_t = y\}$ is the first hitting time of barrier $y \geq 0$, and
- $\mathcal{T}_0$ is the set of all stopping times θ (with respect to X) such that $\theta \leq T_0$ almost surely, i.e., all stopping times until the process X reaches 0.

Optimal stopping problems (see [249]) are stochastic control problems where the control variable is the time, i.e., the controller needs to choose the most appropriate time to perform an action. For example, in quantitative finance, the pricing of American Options (see [167, 243]) is an optimal stopping problem. In [97], the optimal stopping problem given by Eqn. (D.13) is the detection of the closest point of X_t to the current maximum $X^*_{T_0}$ until reaching 0, measured by the function l.

Under mild assumptions on the Itô diffusion (for example if X is a CIR process) and with $l(x) = \frac{ax^2}{2}$, $a > 0$, Espinosa and Touzi find in [97] a free barrier $\gamma(x)$ such that the stopping time $T^* = T_0 \wedge \inf_{t \geq 0}\{X^*_t \geq \gamma(X_t)\}$ is optimal for this problem, i.e., we have detected a peak in X_t at

2. The mean-reverting character of this Itô diffusion is given by $\mu(x) \leq 0$, where the drift coefficient "pulls" the nonnegative diffusion towards 0. It is also interesting to notice that the mean-revertion is a condition on the drift alone, independent of the diffusion coefficient. This is explained by the fact that the drift and the diffusion coefficient are responsible for infinitesimal changes of different magnitudes.

time T^*. Moreover, they show that the free barrier γ has two monotone parts; first a decreasing part $\gamma_\downarrow(x)$ and then an increasing part $\gamma_\uparrow(x)$, such that

$$\begin{cases} \gamma'_\downarrow(x) &= \frac{Lg(x,\gamma_\downarrow)}{1-x\frac{S'(x)}{S(x)}-\frac{S(x)}{S(\gamma_\downarrow)}} \\ \gamma_\downarrow(0) &= \inf\{z \geq 0 \mid Lg(0,z) \geq 0\} \\ \\ \gamma'_\uparrow(x) &= \frac{Lg(x,\gamma_\uparrow)}{1-\frac{S(x)}{S(\gamma_\uparrow)}} \\ \gamma_\uparrow(0) &= \inf\{x \geq 0 \mid \gamma'_\downarrow(x) \geq 0\}, \end{cases} \tag{D.14}$$

where (see [97])

- $\alpha(x) = -\frac{2\mu(x)}{\sigma^2(x)}$,
- $S(x) = \int_0^x e^{\int_0^u \alpha(r)dr}du$ satisfies $S''(x) = \alpha(x)S'(x)$ with $S(0) = 0$ and $S'(0) = 1$,
- $g(x,z) = \frac{a}{2}(z-x)^2 + aS(x)\int_z^\infty \frac{u-x}{S(u)}du$ for $0 \leq x \leq z$, and
- $Lv(x,z) = \partial^2_{x^2}v(x,z) - \alpha(x)\partial_x v(x,z)$.

Bibliography

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1974. (Cited in pages 117, 155 and 158.)

[2] D. Acemoglu, A. Ozdaglar, and A. ParandehGheibi. Spread of (mis)information in social networks. *Games and Economic Behavior*, 70(2), 2010. (Cited in page 35.)

[3] E. Adar and L. A. Adamic. Tracking information epidemics in blogspace. *In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 207–214, 2005. (Cited in page 29.)

[4] L. Akoglu, H. Tong, and D. Koutra. Graph based anomaly detection and description: a survey. *Data Mining and Knowledge Discovery*, 29(3):626–688, 2015. (Cited in page 28.)

[5] R. Albert, H. Jeong, and A.-L. Barabási. Diameter of the world wide web. *Nature*, 401, 1999. (Cited in page 18.)

[6] R. Albright, J. Cox, D. Duling, A. N. Langville, and C. D. Meyer. Algorithms, initializations, and convergence for the nonnegative matrix factorization. *SAS Technical report, ArXiv: 1407.7299*, 2014. (Cited in page 98.)

[7] D. Aldous. Interacting particle systems as stochastic social dynamics. *Bernoulli*, 19(4), 2013. (Cited in page 24.)

[8] J. Allan. *Topic Detection and Tracking*, volume 12 of *The Information Retrieval Series*. Kluwer, 2002. (Cited in page 28.)

[9] C. Alós-Ferrer and Nick Netzer. The logit-response dynamics. *Games and Economic Behavior*, 68(2):413–427, 2010. (Cited in page 25.)

[10] L. AlSumait, D. Barbará, and C. Domeniconi. On-line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. *In Proceeding of the 8th IEEE International Conference on Data Mining (ICDM)*, pages 3–12, 2008. (Cited in page 29.)

[11] A. Arenas, A. Fernández, and S Gómez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10, 2008. (Cited in page 20.)

[12] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. *In Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009. (Cited in page 148.)

[13] K. Atkinson. *A Survey of Numerical Methods for the Solution of Fredholm Integral Equations of the Second Kind*. Society for Industrial and Applied Mathematics, 1976. (Cited in page 100.)

[14] K. Avrachenkov, M. El Chamie, and G. Neglia. Graph clustering based on mixing time of random walks. *In proceedings of the IEEE International Conference on Communications (ICC)*, 2014. (Cited in pages 19 and 22.)

[15] E. Bacry, S. Delattre, M. Hoffmann, and J.-F. Muzy. Modelling microstructure noise with mutually exciting point processes. *arXiv:1101.3422v1*, November 2011. (Cited in pages 78, 100 and 138.)

[16] E. Bacry, T. Jaisson, and J.-F. Muzy. Estimation of slowly decreasing hawkes kernels: Application to high frequency order book modelling. *ArXiv: 1412.7096*, 2014. (Cited in page 100.)

[17] E. Bacry and J.-F. Muzy. Second order statistics characterization of Hawkes processes and non-parametric estimation. *ArXiv: 1401.0903v1*, 2014. (Cited in pages 100, 140, 141 and 142.)

[18] E. Bacry and J.-F. Muzy S. Gaïffas. Concentration for matrix martingales in continuous time and microscopic activity of social networks. *ArXiv: 1412.7705*, 2014. (Cited in page 100.)

[19] E. Bacry, S.Gaïffas, and J.-F. Muzy. A generalization error bound for sparse and low-rank multivariate hawkes processes. *ArXiv: 1501.00725*, 2015. (Cited in page 100.)

[20] E. Bakshy, I. Rosenn, Cameron Marlow, and L. Adamic. The role of social networks in information diffusion. *In Proceedings of the 21st International Conference on World Wide Web (WWW)*, pages 519–528, 2012. (Cited in page 30.)

[21] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999. (Cited in pages 18 and 50.)

[22] N. Barbieri, F. Bonchi, and G. Manco. Topic-aware social influence propagation models. *In Proceedings of IEEE 12th International Conference on Data Mining (ICDM)*, 2012. (Cited in page 28.)

[23] M. Benaïm. Dynamics of stochastic approximation algorithms. *Séminaires de probabilités de Strasbourg*, 33:1–68, 1999. (Cited in pages 39, 40, 41, 49 and 135.)

[24] E. Berger. Dynamic monopolies of constant size. *Journal of Combinatorial Theory Series B*, 83, 2001. (Cited in page 24.)

[25] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. *In Proceedings of the 3rd International Conference on Internet and Network Economics (WINE)*, pages 306–311, 2007. (Cited in page 28.)

[26] S. Bikhchandani, D. Hirshleifer, and I. Welch. A theory of fads, fashion, custom and cultural change as information cascades. *Journal of Political Economy*, 100, 1992. (Cited in pages 24 and 77.)

[27] P. Billingsley. *Convergence of probability measures*, volume 493 of *Wiley series in Probability and Statistics*. Wiley, New York, 2009. (Cited in page 116.)

[28] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, 2006. (Cited in pages 25, 33, 36, 38 and 149.)

[29] M. Blatt, S. Wiseman, and E. Domany. Super-paramagnetic clustering of data. *Physical Review Letters*, 76(3251), 1996. (Cited in pages 53 and 127.)

[30] D. Blei and J. Lafferty. Dynamic topic models. *In Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006. (Cited in page 162.)

[31] D. Blei and J. McAuliffe. Supervised topic models. *Neural Information Processing Systems (NIPS)*, 21, 2007. (Cited in page 162.)

[32] D. M. Blei, M. I. Jordan, T. L. Griffiths, and J. B. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in Neural Information Processing Systems (NIPS) 16: Proceedings of the 2003 Conference*, 2003. (Cited in page 162.)

[33] D. M. Blei and J. D. Lafferty. Topic models. *Notes*. (Cited in pages 78 and 82.)

[34] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2009. (Cited in pages 29, 78, 82, 93, 145, 146, 148, 152, 153, 155, 156, 157, 160 and 162.)

[35] V. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, October 2008. (Cited in pages 19, 20 and 64.)

[36] V. D. Blondel, N.-D. Ho, and P. van Dooren. Weighted nonnegative matrix factorization and face feature extraction. *Image and Vision Computing*, pages 1–17, 2007. (Cited in page 95.)

[37] L. Blume. The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 5, 1993. (Cited in page 25.)

[38] C. Blundell, J. Beck, and K. A. Heller. Modelling reciprocating relationships with Hawkes processes. *Advances in Neural Information Processing Systems (NIPS)*, 2012. (Cited in pages 27 and 78.)

[39] B. Bollobás. *Random Graphs*. Cambridge University Press, 2nd edition, 2001. (Cited in pages 18 and 126.)

[40] G. N. Borisyuk, R. M. Borisyuk, A. B. Kirillov, E. I. Kovalenko, and V. I. Kryukov. A new statistical method for identifying interconnections between neuronal network elements. *Biological Cybernetics*, 52:301–306, 1985. (Cited in page 78.)

[41] V. Borkar. *Stochastic approximation: a dynamical systems viewpoint*. Cambridge University Press, 2008. (Cited in page 39.)

[42] C. Boutsidisa and E. Gallopoulos. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41:1350–1362, 2008. (Cited in page 98.)

[43] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6), 2006. (Cited in page 34.)

[44] P. Brémaud and L. Massoulié. Stability of nonlinear Hawkes processes. *The Annals of Probability*, 24(3):1563–1588, 1996. (Cited in pages 99, 111, 112 and 141.)

[45] C. Budak, D. Agrawal, and A. El Abbadi. Structural trend analysis for online social networks. *In Proceedings of the VLDB Endowment*, 4(10):646–656, 2011. (Cited in page 30.)

[46] M. Burger, M. Di Francesco, P. A. Markowich, and M. T. Wolfram. Mean field games with nonlinear mobilities in pedestrian dynamics. *Discrete and Continuous Dynamical Systems*, 19(5):1311–1333, 2014. (Cited in page 18.)

[47] K. R. Canini, L. Shi, and T. L. Griffiths. Online inference of topics with latent Dirichlet allocation. *In Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009. (Cited in pages 148 and 150.)

[48] C. Castellano, S. Fortunato, and V. Loreto. Statistical physics of social dynamics. *Reviews of Modern Physics*, 81, 2009. (Cited in pages 18 and 33.)

[49] C. Castellano, M. A. Muñoz, and R. Pastor-Satorras. Nonlinear q-voter model. *Physical Review E*, 80, 2009. (Cited in page 34.)

[50] M. Cataldi, L. Di Caro, , and C. Schifanella. Emerging topic detection on twitter based on temporal and social terms evaluation. *In Proceeding of the 10th International Workshop on Multimedia Data Mining (MDMKDD)*, 2010. (Cited in pages 30, 109 and 110.)

[51] D. Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996), 2010. (Cited in pages 28, 29 and 77.)

[52] D. Centola and M. Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3), 2007. (Cited in pages 28 and 29.)

[53] R. Cerf and A. Pisztora. On the wulff crystal in the ising model. *Annals of probability*, pages 947–1017, 2000. (Cited in page 33.)

[54] G. A. Chandlen and I. G. Graham. The convergence of Nyström methods for Wiener-Hopf equations. *Numerische Mathematik*, 52, 1988. (Cited in page 142.)

[55] W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincon, X. Sun, Y. Wang, W. Wei, and Y. Yuan. Influence maximization in social networks when negative opinions may emerge and propagate. *In Proceedings of the 11th SIAM International Conference on Data Mining (SDM)*, 2011. (Cited in page 28.)

[56] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. *In Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2009. (Cited in page 28.)

[57] W. Chen, Y. Wang, and S. Yang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. *In Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2010. (Cited in page 28.)

[58] J. Cheng, L. Adamic, A. Dow, J. Kleinberg, and J. Leskovec. Can cascades be predicted? *In Proceedings of ACM International Conference on World Wide Web (WWW)*, 2014. (Cited in pages 28 and 30.)

[59] F. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997. (Cited in page 132.)

[60] V. Chvatal. A greedy heuristic for the set-covering problem. *Mathematics of operations research*, 4(3):233–235, 1979. (Cited in page 26.)

[61] A. Cichocki, S. Amari, R. Zdunek, R. Kompass, G. Hori, and Z. He. Extended smart algorithms for non-negative matrix factorization. *Springer LNAI*, 4029, 2006. (Cited in page 98.)

[62] A. Cichocki and R. Zdunek. Regularized alternating least squares algorithms for non-negative matrix/tensor factorizations. In *Advances in Neural Networks - ISNN 2007*, volume 4493 of *Lecture Notes in Computer Science*, pages 793–802. 2007. (Cited in pages 98 and 143.)

[63] A. Cichocki, R. Zdunek, and S. Amari. Nonnegative matrix and tensor factorization. *IEEE Signal Processing Magazine*, 25(1), 2008. (Cited in page 98.)

[64] A. Cichocki, R. Zdunek, and S.-I. Amari. Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization. *Springer LNCS*, 4666, 2007. (Cited in page 98.)

[65] A. Cichocki, R. Zdunek, S. Choi, R. Plemmons, and S.-I. Amari. Novel multi-layer nonnegative tensor factorization with sparsity constraints. *Springer LNCS*, 4432, 2007. (Cited in page 98.)

[66] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. John Wiley & Sons, Ltd, 2009. (Cited in pages 85 and 92.)

[67] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70:39–43, 2004. (Cited in page 20.)

[68] P. Clifford and A. W. Sudbury. A model for spatial conflict. *Biometrika*, 60:581–588, 1973. (Cited in pages 16 and 34.)

[69] J. C. Cox, J. E. Ingersoll, and S. A. Ross. A theory of the term structure of interest rates. *Econometrica*, 53:385–407, 1985. (Cited in pages 112 and 163.)

[70] J. T. Cox and R. Durrett. Nonlinear voter models. *In Random Walks, Brownian Motion, and Interacting Particle Systems*, pages 189–202, 1991. (Cited in page 16.)

[71] R. Crane and D. Sornette. Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, 105(41):15649–15653, 2008. (Cited in pages 78, 81, 99, 109, 111 and 137.)

[72] F. Cucker and S. Smale. Emergent behavior in flocks. *Automatic Control, IEEE Transactions on*, 52(5):852–862, 2007. (Cited in pages 17 and 33.)

[73] A. Czirók and T. Vicsek. Collective behavior of interacting self-propelled particles. *Physica A*, 281:17–29, 2006. (Cited in page 17.)

[74] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes*. Springer series in Statistics. Springer, 2005. (Cited in pages 26, 78, 79, 85, 86, 111, 112, 137 and 150.)

[75] L. Dall'Asta and C. Castellano. Effective surface-tension in the noise-reduced voter model. *Europhysics Letters*, 77(6), 2007. (Cited in page 34.)

[76] H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schölkopf. Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm. 2014. (Cited in page 27.)

[77] W. M. Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. *Technical report*, 2011. (Cited in pages 78, 148, 150 and 151.)

[78] A. Das, S. Gollapudi, and K. Munagala. Modeling opinion dynamics in social networks. *In Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM)*, 2014. (Cited in page 35.)

[79] R. Dawkins. *The Selfish Gene*. Oxford University Press, 2nd edition, 1989. (Cited in pages 26 and 78.)

[80] K. Al Dayri, E. Bacry, and J.-F. Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *European Physical Journal B*, 85(157), 2012. (Cited in page 100.)

[81] M. S. de la Lama, J. M. López, and H. S. Wio. Spontaneous emergence of contrarian-like behaviour in an opinion spreading mode. *EPL (Europhysics Letters)*, 72(25):851–857, 2005. (Cited in page 17.)

[82] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 3(1–4):87–98, 2000. (Cited in pages 15 and 34.)

[83] M. H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974. (Cited in pages 15 and 34.)

[84] S. Delattre, N. Fournier, and M. Hoffmann. High dimensional hawkes processes. *ArXiv: 1403.5764*, 2014. (Cited in page 103.)

[85] J. P. Desai, V. Kumar, and J. P. Ostrowski. Control of changes in formation for a team of mobile robots. *In Proceedings of IEEE International Conference on Robotics and Automation*, 1999. (Cited in page 33.)

[86] P. Dodds and D. Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21), 2004. (Cited in page 29.)

[87] M. De Domenico, A. Sole-Ribalta, E. Cozzo, M. Kivela, Y. Moreno, M. A. Porter, S. Gómez, and A. Arenas. Mathematical formulation of multi-layer networks. *Physical Review X*, (3), 2013. (Cited in pages 82, 126 and 127.)

[88] P. Domingos and M. Richardson. Mining the network value of customers. *In Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pages 57–66, 2001. (Cited in page 28.)

[89] I. Douven and A. Riegler. Extending the Hegselmann-Krause model iii: From single beliefs to complex belief states. *Episteme*, 6(2):145–163, 2009. (Cited in page 15.)

[90] I. Douven and A. Riegler. Extending the Hegselmann-Krause model i. *The Logic Journal of the IGPL*, 18(2):323–335, 2010. (Cited in page 15.)

[91] I. Douven and A. Riegler. Extending the Hegselmann-Krause model ii. *The analytical way*, pages 245–258, 2010. (Cited in page 15.)

[92] E. J. Dudewicz and S. N. Mishra. *Modern Mathematical Statistics*. Wiley, 1988. (Cited in page 104.)

[93] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press, 2010. (Cited in page 23.)

[94] R. Elie and G.-E. Espinosa. Optimal stopping of a mean reverting diffusion: minimizing the relative distance to the maximum. *hal: 00573429*, 2011. (Cited in page 175.)

[95] G. Ellison. Learning, local interaction, and coordination. *Econometrica*, 61, 1993. (Cited in page 25.)

[96] P. Erdös and A. Rényi. On random graphs I. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959. (Cited in pages 18 and 50.)

[97] G.-E. Espinosa and N. Touzi. Detecting the maximum of a scalar diffusion with negative drift. *SIAM Journal on Control and Optimization*, 50(5):2543–2572, 2012. (Cited in pages 113, 115, 117, 119, 163, 175, 176 and 177.)

[98] E. Estrada and N. Hatano. Communicability in complex networks. *Physical Review E*, 77(3), 2008. (Cited in pages 19 and 23.)

[99] M. Farajtabar, N. Du, M. Gomez-Rodriguez, I. Valera, H. Zha, and L. Song. Shaping social activity by incentivizing users. 2014. (Cited in pages 26 and 109.)

[100] C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011. (Cited in pages 86, 87, 88 and 90.)

[101] M. Fliess. A mathematical proof of the existence of trends in financial time series. *Systems Theory: Modelling, Analysis and Control*, pages 43–62, 2009. (Cited in page 28.)

[102] S. Fortunato. Community detection in graphs. *Physics Reports*, 468(3-5):75–174, 2010. (Cited in pages 19, 20 and 23.)

[103] S. Fortunato and M. Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007. (Cited in pages 20 and 61.)

[104] L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977. (Cited in page 23.)

[105] J. R. P. French. A formal theory of social power. *Psychological review*, 63(3), 1956. (Cited in pages 15 and 33.)

[106] D. Fudenberg and S. Takahashi. Heterogeneous beliefs and local information in stochastic fictitious play. *Games and Economic Behavior*, 71(1), 2011. (Cited in page 126.)

[107] N. Fyson, T. De Bie, and N. Cristianini. Reconstruction of causal networks by set covering. *In Proceedings of the International Conference on Adaptive and Natural Computing Algorithms (ICANNGA)*, 2011. (Cited in page 26.)

[108] S. Galam. Majority rule, hierarchical structures and democratic totalitarism: a statistical approach. *Journal of Mathematical Psychology*, 30:426–434, 1986. (Cited in page 16.)

[109] S. Galam. Minority opinion spreading in random geometry. *European Physics Journal B*, 25(4), 2002. (Cited in page 16.)

[110] S. Galam. Local dynamics vs. social mechanisms: A unifying frame. *EPL (Europhysics Letters)*, 70(6), 2005. (Cited in page 17.)

[111] A. Galeotti, S. Goyal, M. O. Jackson, F. Vega-Redondo, and L. Yariv. Network games. *Review of Economic Studies*, 77(1):218–244, 2010. (Cited in page 25.)

[112] L. Gao, C. Song, Z. Gao, A.-L. Barabási, J. P. Bagrow, and D. Wang. Quantifying information flow during emergencies. *Nature*, 4, 2014. (Cited in page 29.)

[113] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. CRC Press, 2nd edition, 2003. (Cited in page 149.)

[114] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 1984. (Cited in page 149.)

[115] W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Pratice*. Interdisciplinary Statistics. Chapman & Hall/CRC, 1999. (Cited in page 149.)

[116] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002. (Cited in pages 18, 19, 23, 50 and 64.)

[117] D. F. Gleich. Pagerank beyond the web. *ArXiv: 1407.5107*, 2014. (Cited in page 114.)

[118] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM*, 42(6), 1995. (Cited in page 19.)

[119] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 12(3):211–223, 2001. (Cited in pages 24, 28 and 77.)

[120] J. Goldenberg, B. Libai, and E. Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001. (Cited in pages 24, 28, 77 and 109.)

[121] D. A. Gomes, J. Mohr, and R. R. Sousa. Continuous time finite state space mean-field games. *Applied Mathematics and Optimization*, 68(1):99–143, 2013. (Cited in page 18.)

[122] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. *International Conference on Machine Learning (ICML)*, pages 561–568, 2011. (Cited in pages 27 and 28.)

[123] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2012. (Cited in pages 26, 27 and 77.)

[124] M. Gomez-Rodriguez, J. Leskovec, and B. Schoelkopf. Modeling information propagation with survival theory. *International Conference on Machine Learning (ICML)*, 2013. (Cited in pages 27 and 77.)

[125] M. Gomez-Rodriguez and B. Schölkopf. Influence maximization in continuous time diffusion networks. *In Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 313–320, 2012. (Cited in pages 26 and 28.)

[126] M. Gomez-Rodriguez and B. Schölkopf. Submodular inference of diffusion networks from multiple trees. *International Conference on Machine Learning (ICML)*, 2012. (Cited in pages 26 and 77.)

[127] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978. (Cited in page 24.)

[128] T. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. *Technical report*, 2002. (Cited in pages 78, 148 and 150.)

[129] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. *In L. K. Saul, Y. Weiss and L. Bottou, eds., Advances in Neural Information Processing Systems*, 17:537–544, 2005. (Cited in page 162.)

[130] A. Guille and C. Favre. Mention-anomaly-based event detection and tracking in Twitter. *In Proceedings of the IEEE/ACM International Conference on Advances in Social Network Analysis and Mining (ASONAM)*, 2014. (Cited in page 30.)

[131] A. Guille and H. Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. *In Proceedings of the 21st international conference companion on World Wide Web (WWW)*, pages 1145–1152, 2012. (Cited in page 30.)

[132] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *ACM SIGMOD Record*, 42(2):17–28, 2013. (Cited in page 27.)

[133] P. F. Halpin. An EM algorithm for Hawkes process. *Proceedings of the 77th Annual Meeting of the Psychometric Society*. (Cited in page 81.)

[134] B. D. Hankin and R. A. Wright. Passenger flow in subways. *Operational Research*, 9(2):81–88, 1958. (Cited in page 17.)

[135] N. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point-processes. *To appear in Bernoulli*. (Cited in page 100.)

[136] S. J. Hardiman, N. Bercot, and J.-P. Bouchaud. Critical reflexivity in financial markets: a Hawkes process analysis. *arXiv: 1302.1405*, 8(3), 2013. (Cited in page 163.)

[137] T. E. Harris. Nearest-neighbor markov interaction processes on multidimensional lattices. *Advances in Mathematics*, 9:66–89, 1972. (Cited in page 16.)

[138] J. C. Harsanyi and R. Selten. *A General Theory of Equilibrium Selection in Games*. The MIT Press, Cambridge, MA, 1998. (Cited in page 126.)

[139] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer-Verlag, second edition, 2009. (Cited in page 89.)

[140] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58:83–90, 1971. (Cited in pages 14, 27, 78, 79, 110, 111, 112 and 141.)

[141] A. G. Hawkes and D. Oakes. A cluster process representation of a self-exciting point process. *J. Appl. Prob.*, 11:493–503, 1974. (Cited in pages 99, 104, 111 and 112.)

[142] R. Hegselmann and U. Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002. (Cited in pages 15, 33 and 34.)

[143] T. Heimo, J. Kumpula, K. Kaski, and J. Saramaki. Detecting modules in dense weighted networks with the potts method. *Journal of Statistical Mechanics*, 2008. (Cited in page 21.)

[144] G. Heinrich. Parameter estimation for text analysis. *Technical report*, 2009. (Cited in pages 148 and 151.)

[145] G. Heinrich and M. Goesele. Variational bayes for generic topic models. *KI 2009: Advances in Artificial Intelligence Lecture Notes in Computer Science*, 5803:161–168, 2009. (Cited in pages 78, 145, 148, 152, 153 and 156.)

[146] D. Helbing. A mathematical model for the behavior of individuals in a social field. *Journal of Mathematical Sociology*, 19(3):189–219, 1994. (Cited in page 18.)

[147] D. Helbing, I. Farkas, and T. Vicsek. Simulating dynamical features of escape panic. *Nature*, 407, 2000. (Cited in page 18.)

[148] L. F. Henderson. The statistics of crowd fluids. *Nature*, 229:381–383, 1971. (Cited in page 17.)

[149] P. J.-J. Herings and R. Peeters. Homotopy methods to compute equilibria in game theory. *Economic Theory*, 42(1):119–156, January 2010. (Cited in page 126.)

[150] J. Hofbauer and W. H. Sandholm. Evolution in games with randomly disturbed payoffs. *Journal of Economic Theory*, 132:47–69, 2007. (Cited in pages 41, 42, 45, 46 and 48.)

[151] M. D. Hoffman, D. M. Blei, and F. Bach. Online learning for latent dirichlet allocation. *Advances in Neural and Information Processing Systems (NIPS) 13*, 2010. (Cited in pages 78, 148 and 152.)

[152] T. Hofmann. Probabilistic latent semantic indexing. *Proceedings of the 22th Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 1999. (Cited in page 145.)

[153] R. A. Holley and T. M. Liggett. Ergodic theorems for weakly interacting infinite systems and the voter model. *Annals of Probability*, 3(4):643–663, 1975. (Cited in page 16.)

[154] P. Holme and J. Saramäki, (eds.). *Temporal Networks*. Springer, Berlin, 2013. (Cited in pages 78, 100, 111, 126 and 127.)

[155] J. A. Holyst, K. Kacperski, and F. Schweitzer. Social impact models of opinion dynamics. *Annual Review of Computational Physics*, 9:253–273, 2001. (Cited in page 17.)

[156] A. R. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press. 1991. (Cited in page 90.)

[157] B. A. Huberman and L. A. Adamic. Information dynamics in the networked world. pages 371–398, 2004. (Cited in page 29.)

[158] T. Iwata, A. Shah, and Z. Ghahramani. Discovering latent influence in online social activities via shared cascade poisson processes. *In Proceedings of the nternational conference on Knowledge discovery and data mining (KDD)*, 2013. (Cited in pages 27, 78 and 85.)

[159] M. O. Jackson. A survey of models of network formation: Stability and efficiency. In *California Institute of Technology*, 2003. (Cited in page 25.)

[160] M. O. Jackson. *Social and Economic Networks*. Princeton University Press. 2008. (Cited in page 25.)

[161] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes*, volume 288. Springer-Verlag, Berlin, 1987. (Cited in pages 171 and 173.)

[162] T. Jaisson and M. Rosenbaum. Limit theorems for nearly unstable hawkes processes. *ArXiv: 1310.2033*, 2013. (Cited in pages 112, 113, 116 and 165.)

[163] T. Jaisson and M. Rosenbaum. Limit theorems for nearly unstable hawkes processes. *The Annals of Applied Probability*, 25(2):600–631, 2015. (Cited in pages 116, 163, 164, 166, 170, 174 and 175.)

[164] A. Janecek and Y. Tan. Iterative improvement of the multiplicative update nmf algorithm using nature-inspired optimization. *In proceedings of 7th International Conference on Natural Computation (ICNC)*, 3, 2011. (Cited in page 98.)

[165] A. Janecek and Y. Tan. Swarm intelligence for non-negative matrix factorization. *International Journal of Swarm Intelligence Research*, 2(4), 2011. (Cited in page 98.)

[166] A. Janecek and Y. Tan. Using population based algorithms for initializing nonnegative matrix factorization. *Advances in Swarm Intelligence*, pages 307–316, 2011. (Cited in page 98.)

[167] I. Karatzas and S. E. Shreve. *Methods of Mathematical Finance*. Number 39 in Stochastic Modelling and Applied Probability. Springer, 1998. (Cited in page 176.)

[168] L. Katz. A new status index derived from sociometric index. *Psychometrika*, pages 39–43, 1953. (Cited in page 19.)

[169] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003. (Cited in pages 24, 25, 28, 77 and 109.)

[170] Y.-D. Kim and S. Choi. Nonnegative tucker decomposition. *Proceedings of 25th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007. (Cited in pages 85, 86, 89, 92, 142 and 143.)

[171] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598), 1983. (Cited in page 98.)

[172] M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter. Multilayer networks. *Journal of Complex Networks*, (2), 2014. (Cited in pages 82, 111, 126 and 127.)

[173] J. Kleinberg. Bursty and hierarchical structure in streams. *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101, 2002. (Cited in pages 28, 29, 30, 109 and 110.)

[174] J. Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. *In Algorithmic Game Theory (N. Nisan, T. Roughgarden, E. Tardos, V. Vazirani, eds.)*, 2007. (Cited in page 77.)

[175] G. A. Kohring. Ising models of social impact: the role of cumulative advantage. *Journal de Physique I France*, 6(2):301–308, 1996. (Cited in page 17.)

[176] R. Kompass. A generalized divergence measure fon nonnegative matrix factorization. *Neural Computation*, 19(3):780–791, 2007. (Cited in pages 86, 87 and 88.)

[177] V. Kovala and R. Schwabe. A law of the iterated logarithm for stochastic approximation procedures in d-dimensional euclidean space. *Stochastic Processes and their Applications*, 105:299–313, 2003. (Cited in page 61.)

[178] P. L. Krapivsky, S. Redner, and D. Volovik. Reinforcement-driven spread of innovations and fads. *Journal of Statistical Mechanics Theory and Experiment*, 2011. (Cited in page 34.)

[179] U. Krause. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in Difference Equations*, pages 227–236, 2000. (Cited in page 15.)

[180] G. E. Kreindler and H. P. Young. Rapid innovation diffusion in social networks. *Proceedings of the National Academy of Sciences*, 2014. (Cited in page 35.)

[181] B. Krishnamurthy and J. Wang. On network-aware clustering of web clients. *ACM SIGCOMM Computer Communication Review*, 30(4):97–110, 2000. (Cited in page 19.)

[182] V. Krishnamurthy and A. d'Aspremont. Convex algorithms for nonnegative matrix factorization. *ArXiv: 1207.0318*, May 2007. (Cited in pages 78, 89 and 98.)

[183] J. M. Kumpula, J. Saramaki, K. Kaski, and J. Kertesz. Resolution limit in complex network community detection with potts model approach. *European Physics Journal B*, 56(41), 2007. (Cited in page 20.)

[184] T. G. Kurtz and P. Protter. Weak limit theorems for stochastic integrals and stochastic differential equations. *The Annals of Probability*, 19(3):1035–1070, 1991. (Cited in pages 165 and 175.)

[185] H.J. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Number 35 in Applications of Mathematics. Springer, New York, 2nd edition edition, 2003. (Cited in pages 39 and 41.)

[186] A. Kuznetsov. *Solvable Markov processes*. PhD thesis, University of Toronto, 2004. (Cited in page 117.)

[187] N. Lanchier and C. Neuhauser. Voter model and biased voter model in heterogeneous environments. *Journal of Applied Probability*, 44:770–787, 2007. (Cited in page 16.)

[188] A. Lancichinetti, S. Fortunato, and F. Radicchi. Benchmark graphs for testing community detection algorithms. *Physical Review E*, 78, 2008. (Cited in page 65.)

[189] J.-M. Lasry and P.-L. Lions. Mean field games. *Japanese Journal of Mathematics*, 2, 2007. (Cited in page 18.)

[190] B. Latané. The psychology of social impact. *American Psychologist*, 36(4):343–356, 1981. (Cited in page 17.)

[191] E. Le Martelot and C. Hankin. Fast multi-scale detection of relevant communities in large scale networks. *The Computer Journal, Oxford University Press*, 56(9), 2013. (Cited in pages 64 and 65.)

[192] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. (Cited in pages 85, 86 and 89.)

[193] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001. (Cited in pages 86, 88 and 90.)

[194] Marc Lelarge. Diffusion and cascading behavior in random networks. *Games and Economic Behavior*, 75(2), 2012. (Cited in page 25.)

[195] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506, 2009. (Cited in page 26.)

[196] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, 1(1), 2007. (Cited in page 64.)

[197] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. *SIAM International Conference on Data Mining (SDM)*, 2007. (Cited in page 26.)

[198] E. Lewis and G. O. Mohler. A nonparametric EM algorithm for multiscale Hawkes processes. *Journal of Nonparametric Statistics*, pages 1–16, 2011. (Cited in pages 81, 99, 100 and 138.)

[199] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. (Cited in page 162.)

[200] L. Li, H. Deng, A. Dong, Y. Chang, and H. Zha. Identifying and labeling search tasks via query-based Hawkes processes. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*, pages 731–740, 2014. (Cited in page 148.)

[201] L. Li and H. Zha. Dyadic event attribution in social networks with mixtures of Hawkes processes. *Proceedings of 22nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2013. (Cited in pages 26, 78, 109 and 111.)

[202] L. Li and H. Zha. Learning parametric models for social infectivity in multi-dimensional Hawkes processes. *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI)*, 2014. (Cited in page 27.)

[203] T. M. Liggett. *Interacting Particle Systems*. Springer, 1985. (Cited in page 24.)

[204] T. M. Liggett. Coexistence in threshold voter models. *Annals of Probability*, 22:764–802, 1994. (Cited in page 16.)

[205] T. M. Liggett. *Stochastic Interacting Systems: Contact, Voter and Exclusion Processes*. Springer-Verlag, 1999. (Cited in page 34.)

[206] F. Lillo and J. D. Farmer. The long memory of the efficient market. *Studies in Nonlinear Dynamics & Econometrics*, 8(3), 2004. (Cited in page 163.)

[207] C.-J. Lin. Projected gradient methods for nonnegative matrix factorization. *Neural Computation*, 19(10), 2007. (Cited in page 98.)

[208] T. Liniger. Multivariate Hawkes processes. *ETH Doctoral Dissertation*, (18403), 2009. (Cited in pages 14, 27, 78, 79 and 110.)

[209] J. S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427), 1994. (Cited in page 149.)

[210] J. Lorenz. A stabilization theorem for dynamics of continuous opinions. *Physica A: Statistical Mechanics and its Applications*, 355(1):217–223, 2005. (Cited in page 15.)

[211] J. C. Louzada Pinto and T. Chahed. Modeling multi-topic information diffusion in social networks using latent Dirichlet allocation and Hawkes processes. *The 10th International Conference on Signal Image Technology and Internet Based Systems (SITIS '14)*, 2014. (Cited in pages 82, 85 and 95.)

[212] J. C. Louzada Pinto and T. Chahed. Modeling user and topic interactions in social networks using Hawkes processes. *8th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS '14)*, 2014. (Cited in pages 85, 89 and 90.)

[213] L. Lovász. Random walks on graphs: A survey. 1993. (Cited in page 21.)

[214] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, (54), 2003. (Cited in pages 50 and 64.)

[215] R. P. Malhame M. Y. Huang and P. E. Caines. Large population stochastic dynamic games: Closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Special issue in honor of the 65th birthday of Tyrone Duncan, Communications in Information and Systems*, 6(3):221–252, 2006. (Cited in page 18.)

[216] M. Macy and R. Willer. From factors to actors: Computational sociology and agent-based modeling. *Ann. Rev. Soc.*, 2002. (Cited in pages 24 and 28.)

[217] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 2nd edition, 1999. (Cited in page 21.)

[218] F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95–142, 2013. (Cited in pages 19 and 23.)

[219] X.-L. Mao, Z.-Y. Ming, T.-S. Chua, S. Li, H. Yan, and X. Li. Sshlda: a semi-supervised hierarchical topic model. *In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2012. (Cited in page 162.)

[220] J. Marot and S. Bourennane. Fast tensor signal filtering using fixed point algorithm. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008. (Cited in page 98.)

[221] D. Marsan and O. Lengline. Extending earthquakes' reach through cascading. *Science*, 319(1076), 2008. (Cited in pages 100 and 138.)

[222] M. Masseroli, D. Chicco, and P. Pinoli. Probabilistic latent semantic analysis for prediction of gene ontology annotations. *In proceedings of IEEE World Congress on Computational Intelligence (WCCI)*, 2012. (Cited in page 145.)

[223] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. *In Proceedings of Neural Information Processing Systems (NIPS)*, 2012. (Cited in page 64.)

[224] D. McFadden. Quantal choice analysis: a survey. *Annals of Economic and Social Measurament*, 5(4), 1976. (Cited in page 25.)

[225] R. D. McKelvey and T. R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10:6–38, 1995. (Cited in page 25.)

[226] B. A. Miller, N. T. Bliss, and P. J. Wolfe. Subgraph detection using eigenvector l1 norms. *Advances in Neural Information Processing Systems (NIPS)*, pages 1633–1641, 2010. (Cited in page 21.)

[227] H. Minc. *Nonnegative Matrices*. John Wiley & Sons, New York, NY, USA, 1988. (Cited in page 89.)

[228] T. P. Minka. Estimating a dirichlet distribution. *Notes*, 2012. (Cited in page 153.)

[229] A. Montanari and A. Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47), 2010. (Cited in page 25.)

[230] A. Murua, L. Stanberry, and W. Stuetzle. On potts model clustering, kernel k-means and density estimation. *Journal of Computation and Graphical Statistics*, pages 629–658, 2008. (Cited in page 127.)

[231] S. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. *IEEE International Conference On Data Mining (ICDM)*, 2012. (Cited in pages 26 and 77.)

[232] S. Myers, J. Leskovec, and C. Zhu. Information diffusion and external influence in networks. *KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012. (Cited in pages 26 and 78.)

[233] M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 2004. (Cited in pages 19, 20 and 65.)

[234] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3), 2006. (Cited in pages 19, 20, 109 and 114.)

[235] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8557–8582, 2006. (Cited in pages 19 and 20.)

[236] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 2004. (Cited in pages 19, 21, 64 and 65.)

[237] B. Noble. *Methods based on the Wiener-Hopf technique for the solution of partial differential equations*. Pergamon, 1958. (Cited in page 100.)

[238] R. J. Norris. *Markov chains*. Cambridge University Press, 1998. (Cited in page 132.)

[239] A. Nowak, J. Szamrej, and B. Latané. From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, 97(3):362–376, 1990. (Cited in page 17.)

[240] Y. Ogata. The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Annals of the Institute of Statistical Mathematics*, 30:Part A:243–261, 1978. (Cited in pages 85, 86, 104, 137 and 150.)

[241] Y. Ogata. On lewis simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981. (Cited in pages 103, 111 and 119.)

[242] Y. Ogata. Space-time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics*, 50(2):379–402, 1998. (Cited in page 78.)

[243] B. Oksendal and A. Sulem. *Applied Stochastic Control of Jump Diffusions*. Universitext, 2007. (Cited in page 176.)

[244] B. K. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, Berlin, 2003. (Cited in page 176.)

[245] S. A. Pasha and V. Solo. Hawkes-laguerre dynamic index models for point processes. *Proceedings of 52nd IEEE Conference on Decision and Control (CDC)*, 2013. (Cited in pages 81, 85, 89, 99 and 137.)

[246] S. A. Pasha and V. Solo. Hawkes-laguerre reduced rank model for point process. *Proceedings of 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6098–6102, 2013. (Cited in pages 81, 85, 99 and 137.)

[247] D. Peleg. Local majority voting, small coalitions, and controlling monopolies in graphs: A review. *3rd Colloquium on Structural Information and Communication*, 1996. (Cited in page 24.)

[248] M. Pelletier. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *The Annals of Applied Probability*, 8(1):10–44, 1998. (Cited in pages 61 and 62.)

[249] G. Peskir and A. Shiryaev. *Optimal Stopping and Free-Boundary Problems*. Lectures in Mathematics. ETH Zürich, 2006. (Cited in page 176.)

[250] A.-H. Phan and A. Cichocki. Extended hals algorithm for nonnegative tucker decomposition and its applications for multi-way analysis and classification. *Neurocomputing*, 74:1956–1969, 2011. (Cited in pages 98 and 143.)

[251] A.-H. Phan, A. Cichocki, R. Zdunek, and T. Vu-Dinh. Novel alternating least squares algorithms for nonnegative matrix and tensor factorizations. 6443:262–269, 2010. (Cited in pages 98 and 143.)

[252] M. Plantié and M. Crampes. Survey on social community detection. In N. Ramzan, R. van Zwol, J.-S. Lee, K. Clüver, and X.-S. Hua, editors, *Social Media Retrieval*, Computer Communications and Networks, pages 65–85. 2013. (Cited in page 23.)

[253] P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(191), 2006. (Cited in pages 19 and 22.)

[254] R. B. Potts. *The Mathematical Investigation of Some Cooperative Phenomena*. PhD thesis, University of Oxford, 1951. (Cited in page 33.)

[255] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Finding community structure in networks using the eigenvectors of matrices. *Proceedings of the National Academy of Sciences*, 101(9):2658–2663, 2004. (Cited in pages 57, 66 and 70.)

[256] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3), 2007. (Cited in pages 22, 53, 54, 57, 60 and 64.)

[257] H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, 1961. (Cited in page 149.)

[258] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555, 2002. (Cited in page 19.)

[259] K. P. Reddy, M. Kitsuregawa, P. Sreekanth, and S. S. Rao. A graph based approach to extract a neighborhood customer community for collaborative filtering. *In Proceedings of the Second International Workshop on Databases in Networked Information Systems (DNIS)*, pages 188–200, 2002. (Cited in page 19.)

[260] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74, 2006. (Cited in pages 19, 21, 64, 65 and 68.)

[261] P. Reynaud-Bouret, V. Rivoirard, F. Grammont, and C. Tuleau-Malot. Goodness-of-fit tests and nonparametric adaptive estimation for spike train analysis. *Journal of Mathematical Neuroscience*, 4(3), 2014. (Cited in page 100.)

[262] P. Reynaud-Bouret and S. Schbath. Adaptive estimation for hawkes processes; application to genome analysis. *Annals of Statistics*, 38, 2010. (Cited in page 100.)

[263] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. *In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002. (Cited in page 24.)

[264] E. Rogers. *Diffusion of Innovations*. Free Press, fourth edition, 1995. (Cited in page 24.)

[265] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI '04)*, 2004. (Cited in pages 78, 82, 93, 145, 146, 151 and 152.)

[266] K. Saito, M. Kimura, K. Ohara, and H. Motoda. Selecting information diffusion models over social networks for behavioral analysis. 6323:180–195, 2010. (Cited in page 30.)

[267] S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007. (Cited in page 19.)

[268] T. Schelling. *Micromotives and Macrobehavior*. Norton, 1978. (Cited in page 24.)

[269] F. Schweitzer. *Brownian Agents and Active Particles*. Springer Series in Synergetics. Springer Verlag, Berlin-Heidelberg, Germany, 2003. (Cited in page 17.)

[270] F. Schweitzer and J. Holyst. Modelling collective opinion formation by means of active brownian particles. *European Physical Journal B*, 15(4):723–732, 2000. (Cited in page 17.)

[271] L. Shu, B. Long, and W. Meng. A latent topic model for complete entity resolution. *25th IEEE International Conference on Data Engineering (ICDE)*, 2009. (Cited in pages 146 and 162.)

[272] P. Smaragdis. Convolutive speech bases and their application to supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(1), 2006. (Cited in page 101.)

[273] T. M. Snowsill, N. Fyson, T. De Bie, and N. Cristianini. Refining causality: who copied from whom? *ACM SIGKDD 2011*, pages 466–474, 2011. (Cited in pages 26 and 77.)

[274] K. Sznajd-Weron. Mean-field results for the two-component model. *Physical Review E*, 71(4), 2005. (Cited in page 17.)

[275] K. Sznajd-Weron and J. Sznajd. Opinion evolution in closed community. *International Journal of Modern Physics C*, 11(6), 2000. (Cited in pages 16 and 34.)

[276] T. Takahashi, R. Tomioka, and K. Yamanishi. Discovering emerging topics in social streams via link anomaly detection. *In Proceeding of the 11th IEEE International Conference on Data Mining (ICDM)*, 2011. (Cited in pages 30, 109 and 110.)

[277] V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the β-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1592–1605, 2013. (Cited in page 98.)

[278] L. Tang, X. Wang, and H. Liu. Uncovering groups via heterogeneous interaction analysis. *In Proceedings of IEEE International Conference on Data Mining (ICDM)*, 2009. (Cited in page 64.)

[279] Y. Tang, X. Xiao, and Y. Shi. Influence maximization: near-optimal time complexity meets practical efficiency. *In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD)*, pages 75–86, 2014. (Cited in page 28.)

[280] Y. W. Teh and M. I. Jordan. Hierarchical bayesian nonparametric models with applications. *Bayesian Nonparametrics (Cambridge University Press)*, 2010. (Cited in pages 103, 146 and 162.)

[281] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581, 2006. (Cited in pages 103 and 162.)

[282] C.-Y. Teng, L. Gong, A. L. Eecs, C. Brunetti, and L. Adamic. Coevolution of network structure and content. *In Proceedings of the 4th Annual ACM Web Science Conference (WebSci)*, pages 288–297, 2012. (Cited in page 30.)

[283] G. Tibély and J. Kertész. On the equivalence of the label propagation method of community detection and a Potts model approach. *Physica A*, 387:4982–4984, 2008. (Cited in pages 22, 54, 57 and 60.)

[284] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Royal. Statist. Soc B.*, 58(1), 1996. (Cited in page 89.)

[285] P. Tino. Equilibria of iterative softmax and critical temperatures for intermittent search in self-organizing neural networks. *Neural Computation*, 19(4):1056–1081, 2007. (Cited in pages 59 and 126.)

[286] P. Tino. Bifurcation structure of equilibria of iterated softmax. *Chaos, Solitons & Fractals*, 41:1084–1816, 2009. (Cited in pages 59 and 126.)

[287] V. A. Traag, P. Van Dooren, and Y. Nesterov. Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84:016114, 2011. (Cited in page 61.)

[288] N. Tremblay and P. Borgnat. Multiscale community mining in networks using spectral graph wavelets. *In proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2013. (Cited in page 21.)

[289] J. Tsitsiklis. Problems in decentralized decision making and computation. *Ph. D. Thesis*, 1984. (Cited in page 34.)

[290] T. L. Turocy. A dynamic homotopy interpretation of the logistic quantal response equilibrium correspondence. *Games and Economic Behavior*, 51:243–263, 2005. (Cited in page 126.)

[291] Y. Urabe, K. Yamanishi, R. Tomioka, and H. Iwai. Realtime change-point detection using sequentially discounting normalized maximum likelihood coding. *In Proceedings of the 15th Pacific Asia Knowledge Discovery and Data Mining (PAKDD)*, 2011. (Cited in page 30.)

[292] I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Modeling adoption and usage frequency of competing products and conventions in social media. *Workshop in "Networks: From Graphs to Rich Data" at Neural Information Processing Systems Conference (NIPS)*, 2014. (Cited in page 111.)

[293] S. van Dongen. *Graph Clustering by Flow Simulation*. Phd thesis, University of Utrecht, May 2000. (Cited in pages 21 and 53.)

[294] A. Vandaele, N. Gillis, F. Glineur, and D. Tuyttens. Heuristics for exact nonnegative matrix factorization. *ArXiv: 1411.7245*, 2014. (Cited in page 98.)

[295] G. Vichniac. Simulating physics with cellular automata. *Physica D*, 10:96–115, 1984. (Cited in page 24.)

[296] H. M. Wallach, D. Mimno, and A. McCallum. Rethinking lda: Why priors matter. *In Proceedings of the 23rd Annual Conference on Neural Information Processing Systems (NIPS)*, 2009. (Cited in page 148.)

[297] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. *In Uncertainty in Artificial Intelligence (UAI)*, 2008. (Cited in page 162.)

[298] S. Wang and R. H. Swendsen. Cluster monte carlo algorithms. *Physica A*, 167(565), 1990. (Cited in page 127.)

[299] X. Wang and E. Grimson. Spatial latent dirichlet allocation. *Proceedings of Neural Information Processing Systems Conference (NIPS)*, 2007. (Cited in pages 146 and 162.)

[300] X. Wang, C. Zhai, and R. Sproat X. Hu. Mining correlated bursty topic patterns from coordinated text streams. *In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007. (Cited in pages 28, 29, 109 and 110.)

[301] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 1963. (Cited in page 22.)

[302] D. Watts. A simple model of global cascades in random networks. *Proceedings of the National Academy of Sciences*, 99:5766–5771, 2002. (Cited in pages 24 and 28.)

[303] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini. The role of information diffusion in the evolution of social networks. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 356–364, 2013. (Cited in pages 29 and 30.)

[304] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. *In SIAM International Conference on Data Mining (SDM)*, 2005. (Cited in page 20.)

[305] S. Wild, J. Curry, and A. Dougherty. Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 37, 2004. (Cited in page 98.)

[306] F. Wu and B. A. Huberman. Novelty and collective attention. *Proceedings of the National Academy of Sciences*, 104, 2007. (Cited in page 29.)

[307] F. Y. Wu. The potts model. *Reviews of Modern Physics*, 54(1), 1982. (Cited in pages 21, 22 and 127.)

[308] Q. Wu, L. Zhang, and A. Cichocki. Multifactor sparse feature extraction using convolutive nonnegative tucker decomposition. *Neurocomputing*, 129, 2014. (Cited in page 101.)

[309] Z. Xu, V. Tresp, K. Yu, and H.-P. Kriegel. Infinite hidden relational models. *Uncertainty in Artificial Intelligence (UAI)*, 2006. (Cited in page 27.)

[310] S.-H. Yang and H. Zha. Mixture of mutually exciting processes for viral diffusion. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013. (Cited in pages 27, 78, 81, 85, 99, 103, 104, 107, 109, 111, 137 and 148.)

[311] E. Yildiz, D. Acemoglu, A. Ozdaglar, A. Saberi, and A. Scaglione. Discrete opinion dynamics with stubborn agents. *ACM Transactions on Economics and Computation*, 2012. (Cited in page 16.)

[312] H. P. Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993. (Cited in page 25.)

[313] H. P. Young. The dynamics of social innovation. *Proc Natl Acad Sci USA*, 108(Suppl. 4):21285–21291, 2011. (Cited in pages 18 and 25.)

[314] W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977. (Cited in pages 50 and 64.)

[315] R. Zdunek and A. Cichocki. Nonnegative matrix factorization with constrained second-order optimization. *Signal Processing*, 87(8), 2007. (Cited in page 98.)

[316] K. Zhou, H. Zha, and L. Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013. (Cited in pages 26 and 27.)

[317] K. Zhou, H. Zha, and L. Song. Learning triggering kernels for multi-dimensional Hawkes processes. *Proceedings of the 30th International Conference on Machine Learning (ICML), 2013*, 2013. (Cited in pages 26, 27 and 85.)

[318] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun. Influence maximization in dynamic social networks. *In Proceedings of the IEEE 13th International Conference on Data Mining (ICDM)*, 2013. (Cited in page 28.)

[319] J. Zhuang, Y. Ogata, and D. Vere-Jones. Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97, 2002. (Cited in pages 100 and 138.)